Scientometrics (2022) 127:6011–6028
https://doi.org/10.1007/s11192-022-04484-6

A comparative analysis of local similarity metrics and machine learning approaches: application to link prediction in author citation networks

Adilson Vital[1] · Diego R. Amancio[1]

Received: 21 March 2022 / Accepted: 1 August 2022 / Published online: 12 August 2022

Abstract

Understanding the evolution of paper and author citations is of paramount importance for the design of research policies and evaluation criteria that can promote and accelerate scientific discoveries. Recently many studies on the evolution of science have been conducted in the context of the emergent *Science of Science* field. While many studies have probed the link problem in citation networks, only a few works have analyzed the temporal nature of link prediction in author citation networks. In this study we compared the performance of 10 well-known local network similarity measurements with four machine learning models to predict future links in author citations networks. Differently from traditional link prediction methods, the temporal nature of the predict links is relevant for our approach. Our analysis revealed that the Jaccard coefficient was found to be among the most relevant measurements. The preferential attachment measurement, conversely, displayed the worst performance. We also found that the extension of local measurements to their weighted version do not significantly improved the performance of predicting citations. Finally, we also found that a XGBoost and neural network approach summarizing the information from all 10 considered similarity measurements was able to provide the highest AUC performance and competitive precision values.

Keywords Link prediction · Citation networks · Network similarity · Science of science · Authors citation networks

Introduction

Understanding citation patterns is of paramount importance to understand how science evolves (Fortunato et al. 2018; Nielsen and Andersen 2021). Many efforts have been devoted to understand the mechanisms behind citations (Molléri et al. 2018; Liu et al. 2021). This type of knowledge has allowed an enhanced quantification of evaluation indexes in the *Scientometrics* field (Bai et al. 2016). At the macroscopic level,

✉ Diego R. Amancio
diego@icmc.usp.br

[1] Institute of Mathematics and Computer Science, Department of Computer Science, University of São Paulo, São Carlos, São Paulo, Brazil

paper citations are known to be dependent on age, field, journals visibility and other factors (Amancio et al. 2012b; Krumov et al. 2011). It is well known that citations also are affected by the preferential attachment rule, since more cited papers tend to accrue even more citations. This effect holds for both papers and authors citations (Eom and Fortunato 2011; Wang et al. 2008; Silva et al. 2020).

Several studies have been devoted to understand the mechanisms underlying citations, but most of them have been limited to analyzing and predicting citation counts (Silva et al. 2020). Amancio et al. (2012b) proposed a model that considers three features to predict the behavior of papers citation and authors' h-index. The model considered the preferential attachment rule, the semantical similarity between papers and a memory effect to mimic the tendency of older papers being less cited. While this and other models have been effective to reproduce the *distribution of citations* (and other network measurements), they did not assess the actual correspondence of each *individual edge*. This means that microscopic citation behavior might not be reproduced even though macroscopic features are consistent with the behavior of real-world citation networks. Similar citation distribution analyses have also been performed at the author level (Silva et al. 2020).

A more detailed citation analysis considering both end points of a citation can be performed via link prediction techniques (Lü and Zhou 2011). A comparison of similarity measurements was performed in the context of predicting links in paper citation networks (Shibata et al. 2012). The authors found that the Jaccard coefficient and betweenness centrality affect the predictability of the machine learning system. In addition, a dependency on how the fields are organized was reported, since most predictive systems predicts citations within the same field. Temporal link prediction has also been studied in patent citation networks (Chen et al. 2019). Surprisingly, Chen et al. (2019) found that when mapping local similarities and capturing global structure information, structural deep network embedding is not a good measurement for the task of predicting citations between patents.

While most works in predicting future citations have been performed at the paper/document level, here we focus on predicting citations between authors. Studying the individual citation behavior of particular interest because it can reveal the emergence of individual citation patterns (Radicchi et al. 2009; Fortunato et al. 2018). This type of information can be used not only for evaluation purposes, but can be used to understand how a field evolves (Silva et al. 2016; Powell et al. 2005). Because most citation behavior implies some type of similarity between authors, predicting author citations could also be used to suggest potential effective collaborations (Lande et al. 2020).

In the context of predicting authors citations, here we carried out a comparison of traditional local network similarity measurements for the task of predicting citations between authors. We conducted our link prediction comparative analysis in a dataset comprising more than 450, 000 papers published in Physics journals. Differently from other studies based on author analysis, our methodology is not impacted by authors' names ambiguity (Zhang and Ban 2020; Sebo et al. 2021; Milojević 2013; Amancio et al. 2015; Nie et al. 2021). The considered dataset is enriched with names information extracted from the *Microsoft Academic Graph* (Hug and Brändle 2017).

We limited our comparative analysis to local traditional network measurements for two main reasons: (i) local network measurements can be efficiently computed in very large datasets, with good accuracy results (Martinčić-Ipšić et al. 2017). (ii) the same idea of local neighborhood analysis can be extended to include further hierarchies. Thus, quasi-local similarity measurements can be introduced using the same local measurements (Amancio et al. 2015). Our analysis considered local network similarity measurements and their respective definition in weighted networks. Owing to the popularity of machine learning strategies in a myriad of applications, we also evaluated the effectiveness of four machine learning models, like SVM, Logistic Regression, Artificial Neural Networks and XGBoost (Amancio et al. 2014), combining evidence from all the considered similarity measurements.

Several interesting results were observed in our comparative analysis. All local measurements were found to yield a better precision performance when links are evaluated in a longer time window. The number of citations established between authors did not improved the predictability of citations, since the performance observed with unweighted indexes and their respective weighted versions turned out to be similar in several cases. All in all, the best performance was achieved with the Jaccard coefficient. We also found that combining all similarity network measurements via machine learning does not improve the prediction accuracy. Finally, we also found that the preferential attachment rule should be used in combination with other approaches, since this measurement alone turned out to display a low predictive power in author citation networks.

Methodology

This section presents the methodology used in this study. Section 2.1 describes the dataset used to analyze authors citations. Section 2.2 details the construction of author citation networks. The measurements used to quantify the similarity between two authors are described in Sect. 2.3. Machine learning approaches to address the link prediction task are described in Sect. 2.4. Finally, we report our comparative analysis in Sect. 3. Perspectives for future works are presented in Sect. 4.

Dataset

We used the dataset of papers provided by the *American Physical Society* (APS), which comprises about 450, 000 articles from several APS journals, including *Physical Review Letters*, *Physical Review A–E* and *Reviews of Modern Physics*. This dataset has been largely used in several other studies (Bai et al. 2020; Chacon et al. 2020; Li et al. 2019; Silva et al. 2020). Citations and additional paper metadata are also included in the APS dataset. Examples of metadata are paper DOI, journal name, title, list of authors, affiliations and PACS code.

In order to avoid noise from names ambiguity, we used Microsoft Academic Graph (MAG) information to obtain authors' names. This same procedure has been used in related works analyzing author citation networks (Silva et al. 2020), Because MAG provides a unique identifier for each author, we also avoid the name split issue, i.e. when a single author appear with different names in different publications. In sum, while citations at the paper level are obtained from the APS dataset, we used MAG as an additional dataset to address both name ambiguity and name split issues.

The dataset and code used in our experiments is available at this link.

Network construction

Author-citation networks are constructed using the following methodology. Given a time interval, we use information from papers to obtain citation between authors. Two authors *X* and *Y* are connected by a citation in a given time interval if a paper co-authored by *X* cited at least one paper co-authored by Y. In the weighted version of the network, edges weight represents how many times X cited Y.

Figure 1 illustrates the process of creating author-citation networks from paper citations. The figure shows that article 1, co-authored by A and B, cites a paper co-authored by C, D and E. According to this information, the following links between authors are created: $A \to C$, $A \to D$, $A \to E$, $B \to C$, $B \to D$ and $B \to E$. All edges from the toy dataset illustrated in Fig. 1a are depicted in the graph represented in Fig. 1b (continuous edges). Note that not all pairs of authors are linked in Fig. 1b (see blue dashed lines). These are the potential future links that are evaluated in the link prediction task. Here we divided the links in three main groups, as we can see in Fig. 1c. Blue dashed edges for possible future edges and the dark grey for actual links.

Link prediction

Once the network is constructed, our aim is to predict citations between authors. Two frameworks are commonly used for the task (Wang et al. 2014). The first approach is based on nodes similarity. According to this approach, a similarity value is extracted from all possible links and then sorted in decreasing order. Given a threshold, the considered predicted edges are those taking similarity values above the specified threshold. A different approach consists in considering similarity measurements as features in classification systems. Thus, patterns of links creation are obtained based on previous link

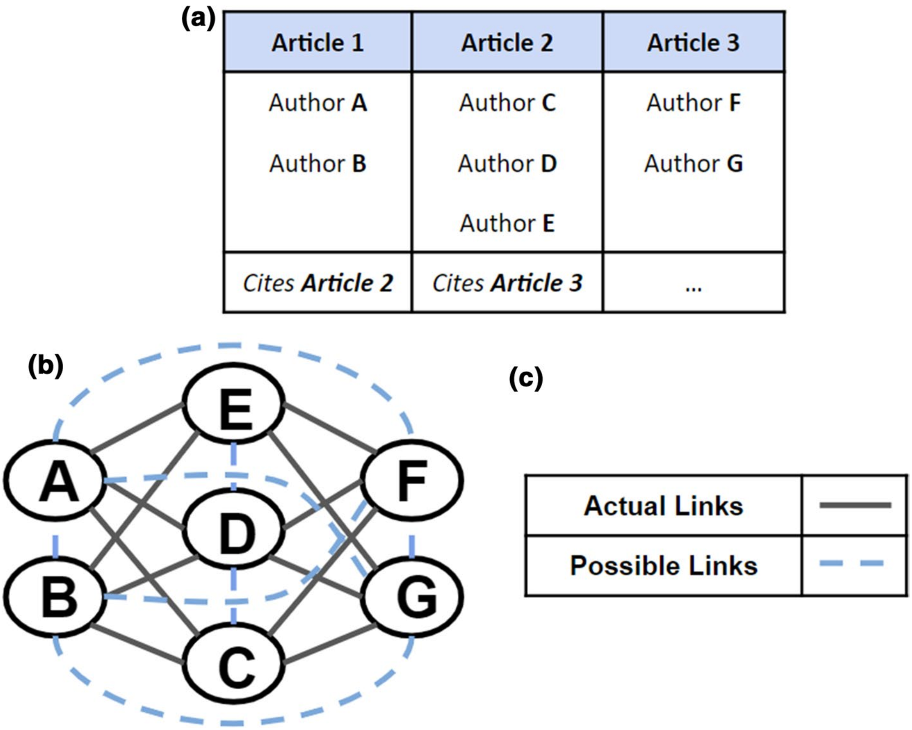

Fig. 1 Representation of the network construction from the articles to the author-citation network

creation dynamics. Here we compared the performance of well-known local network similarity measurements. We also used four machine learning methods to investigate whether simple local measurements are outperformed by an automatic machine learning strategy.

Our analysis was restricted to the 1000 most productive authors observed in the test and training dataset. The reasons for analyzing only the most productive authors are two-fold: (i) productive authors are the ones most active in the field, so it is expected that they are active along many years. This minimizes the issue of trying to predict links between authors that have stopped publishing after a few papers have been published; and (ii) limiting the study to the most productive authors allows us to analyze the citing behavior of many influential researchers who receive a large fraction of citations in the whole author-citation network (Wang et al. 2008).

In the similarity-based strategy, we computed pairwise similarities between all selected authors. Ten different measurements were used. The similarity computation was performed in the training dataset and then the performance of the prediction was evaluated using typical evaluation metrics (see Sect. 2.5 for more details regarding the evaluation). The similarity values were then sorted in decreasing order, and the most similar edges were considered as predicted links according to a threshold value.

In the similarity-based approach, the following similarity measurements were used:

1. *Common Neighbors* (CN): This is one of the simplest and most used similarity measurements (Newman 2001). It quantifies the total number of shared neighbors. Alternatively, this measurement can be regarded as the number of paths of length 2 connecting two nodes. Mathematically, the similarity $\mathrm{CN}(u, v)$ between nodes u and v is computed as

$$\mathrm{CN}(u, v) = |\Gamma(u) \cap \Gamma(v)|, \tag{1}$$

where $\Gamma(v)$ is the set comprising the neighbors of v. The weighted version of this measurement, defined in Lü and Zhou (2010), is given by:

$$\mathrm{WCN}(u, v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} (w_{uz} + w_{vz}), \tag{2}$$

where w_{uz} denotes the weight linking nodes u and v.

2. *Jaccard coefficient* (JC): Another widely used similarity technique is the Jaccard coefficient. This index is similar to CN with the advantage of being normalized in relation to the sum of all neighbors connecting the two data nodes under analysis, i.e.:

$$\mathrm{JC}(u, v) = \frac{|\;\Gamma(u) \cap \Gamma(v)\;|}{|\;\Gamma(u) \cup \Gamma(v)\;|}. \tag{3}$$

While in CN two hubs are more likely to share a neighbor just by chance than low-connected nodes, this effect is minimized by the normalization in Eq. 3. The weighted version of the Jaccard Index (de Sá and Prudencio 2011) is given by:

$$\mathrm{WJC}(u, v) = \frac{\sum_{z \in \Gamma(u) \cap \Gamma(v)} (w_{uz} + w_{vz})}{\sum_{a \in \Gamma(u)} w_{au} + \sum_{b \in \Gamma(v)} w_{bv}}. \tag{4}$$

3. *Adamic-Adar* (AA): this measurement quantifies the similarity between nodes u and v based on the degree (i.e. the number of neighbors) of nodes in $\Gamma(u) \cap \Gamma(v)$ (Adamic and Adar 2003). Mathematically, it is defined as:

$$\mathrm{AA}(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(z)|}. \tag{5}$$

Note that nodes in $\Gamma(u) \cap \Gamma(v)$ with higher degrees contribute with a lower weight in the computation of the similarity between u and v. The term in the denominator of Eq. 5 minimizes the contribution of $z \in \Gamma(u) \cap \Gamma(v)$ whenever z is a hub. This is necessary because hubs are more likely to be connected to both u and v just by chance. The weighted version of the Adamic-Adar measurement (Lü and Zhou 2010) is defined as

$$\mathrm{WAA}(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{w_{uz} + w_{vz}}{\log(1 + \sum_{a \in \Gamma(z)} w_{za})}. \tag{6}$$

4. *Resource Allocation* (RA): Similar to the Adamic-Adar technique, the resource allocation (Zhou et al. 2009) similarity index aims to give lower weight for shared neighbors with a higher degree:

$$\mathrm{RA}(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{|\Gamma(z)|}. \tag{7}$$

Notice that here higher degree neighbors contribute with an even lower weight since $\log |\Gamma(z)|$ in Eq. 5 has now been replaced by $|\Gamma(z)|$ in Eq. 7. The weighted version of the RA index also punishes neighbors with high strength (s) (Lü and Zhou 2010):

$$\mathrm{WRA}(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{w_{uz} + w_{vz}}{s_z}. \tag{8}$$

5. *Preferential Attachment* (PA): This similarity metric is proportional to the product of the degree of the nodes u and v being analyzed (Barabási et al. 2002):

$$\mathrm{PA}(u,v) = |\Gamma(u)| \times |\Gamma(v)|. \tag{9}$$

Because the preferential attachment states the higher-degree nodes are more likely to accrue new links (see e.g. Wang et al. 2008), a new link between two highly connected nodes are very likely to appear in the future. The null model used to quantify the modularity measurement also considers that the probability of two nodes being linked when links are randomly placed is proportional to the product of the degrees of nodes at the end of the edge. The weighted version of the measurement considers the in-strength of nodes instead of the number of neighbors (de Sá and Prudencio 2011):

$$\mathrm{WPA}(u,v) = \sum_{a \in \Gamma(u)} w_{au} \times \sum_{b \in \Gamma(v)} w_{bv}. \tag{10}$$

Machine Learning Approach and Hyperparameter Optimization

In this method, citations are predict via learning models. The similarity measures described in Sect. 2.3 were used as features. The edges of the network are used as positive instances, while non-existent links (i.e. the dashed lines in Fig. 1) were used as negative instances.

Four machine learning models were used in our comparative analysis: (i) artificial neural networks (ANN) (Jain et al. 1996), (ii) logistic regression (LR) (Wright 1995), (iii) support vector machines (SVC) (Noble 2006), and extreme gradient boost (XGBoost) (Chen and Guestrin 2016). Apart from ANN, hyperparameters tuning was performed by using the 5-fold cross validation combined with the grid search technique (Refaeilzadeh et al. 2009). For all models, precision and AUC were selected as optimization and satisficing metrics, respectively (Bradley 1997).

The predictions were based on the best configuration of each algorithm, i.e. the configuration with highest precision without overfitting (Yegnanarayana 2009; Nielsen 2015). A description of the considered methods is provided below:

1. *Artificial Neural Networks*: the best configuration of the neural network was formed by one input layer comprising 10 units and one hidden layer, with 16 unites. The last layer comprised one unit. The *sigmoid* function was used in both hidden and output layers (Yegnanarayana 2009; Nielsen 2015). For the training part, we used Adam as optimizer and binary cross entropy as loss function, with a batch size of 32 and 50 epochs of training. The input of the neural network corresponds to the 10 similarity measurements described in Sect. 2.3 and the output is a real number ranging between 0 and 1. In this way, the neural network method can be seen as a a way to combine and summarize all measurements into a single similarity value. We adopted the parameter optimization adopted in (Amancio et al. 2014).
2. *Support Vector Machine*: Support vector machine is a supervised machine learning algorithm that aims to create a hyperplane capable of separate instances from different classes. The output for each instance is a score reflecting the membership score (or probability) to each possible class. For hyperparameter tuning, we used grid search and k-fold cross-validation techniques. In the end, the best configuration was used the RBG kernel and 10^{-4} of kernel coefficient. The best configuration also used 10^{-1} for the regularization parameter (Amancio et al. 2014).
3. *Logistic Regression*: this method multiplies each input value by a weight. The obtained products are summed and the result is used as input for a sigmoid function. The output is a probability of the considered to belong to a speficif class. The best combination of hyperparameters considered L2 as norm. As for the regularization factor, the best value found was 10^{-5}.
4. *Extreme Gradient Boosting Classifier*: XGBoost is an improved version of traditional random forests (Breiman 2001). Differently from random forests, the XGBoost assigns a different weight in the voting process for each generated tree. The most accurate trees are the ones receiving the highest weights. Because the output of this method is a score, calibration is necessary to obtain membership probabilities to each class. The best combination yielded 0.5 and 0.8 for the subsampling ratio of columns and rows, respectively. The best minimum loss reduction required to make a further partition on a leaf (γ) was $\gamma = 0$. The other considered parameters were learning rate = 0.1, maximum depth = 3, regularization term = 10 and ratio of weight control = 1 (Chen and Guestrin 2016).

Evaluation

One of the most traditional means to evaluate the quality of an information retrieval system is to divide the set of edges E into two parts: the training and test edges. The set of training edges will be used to predict the missing edges. A more elaborated method to perform such a division is the k-fold cross validation approach (Kohavi 1995). According to this technique, the dataset of links is separated into k different parts (folds) of preferably equal sizes. $k - 1$ folds are used for training (E_{train}) and the remaining fold

(E_{test}) is used to test the performance of the model. This process must be repeated at k times using different divisions for the test dataset in order to obtain the performance of our prediction model according to the average of the performance over the repetitions. At each moment, a different division is used as test dataset.

Because author-citation networks represent an evolving dynamic system, an evolution of the network structure is expected. Thus, it is natural to expect that new links may appear and old ones may disappear (i.e. no citations between a pair of authors might be observed in the considered period). New nodes can also appear in the network, as new authors are introduced when they publish a paper for the first time. Given that we are focusing on a link prediction task, we are not predicting links involving new nodes, i.e. nodes that were not observed during the training process.

Owing to the temporal nature of the link prediction problem, we used a modification in the evaluation of the system. Given an initial year Y, we consider a past time window of length d and a future time window of length p. Here we aim at predicting links that are formed in our validation database, which consists in all links formed along the interval t_{val}, where $Y \leq t_{\text{val}} \leq Y + p$. In order to train the algorithms, the training dataset uses the information observed along the interval t_{tr}, where $Y - d \leq t_{\text{tr}} < Y$.

In our analysis we varied p so that the prediction quality could be measured at both short- and long-terms. Figure 2 illustrates the division of the dataset when considering $Y = 2016$ as reference year. In the figure, we also considered $d = 4$ years; therefore the train dataset encompasses the years $2012 - 2015$. p varies so that the performance of the model is evaluated for $p = \{0, 1, 2, 3\}$ years after the reference year Y.

Results and discussion

The comparison of performance is divided into two parts. We first analyze the precision in Sect. 3.1. The analysis considering the *receiver operating characteristic* (ROC) curve is then discussed in Sect. 3.2. While the precision evaluates the accuracy of the model in predicting positive links, the ROC analysis also evaluates the accuracy in not predicting absent future links. In Sect. 3.3, we analyze the correlations between the unweighted and weighted versions of the considered similarity metrics.

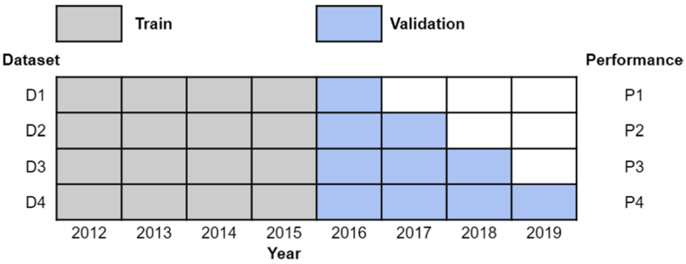

Fig. 2 Illustration of the evaluation methodology used in this analysis. In the first evaluation setting, 2012–2015 is used as train dataset and 2016 is used as validation dataset. Note that the size of the test dataset increases so that one can evaluate both short- and long-term prediction performance

Precision analysis

In Fig. 3, we show the individual behavior of each similarity measurement as more citations are predicted. In our analysis, we considered different validation sets. The first one corresponds to the period between 2000 and 2001. This curve is represented as a blue curve in Fig. 3. The largest validation set corresponds to citations evaluated in the period between 2000 and 2016 (see red curve). Curves ranging from blue to red correspond to test sets considering papers published in the interval [2000, y], where $2001 \leq y \leq 2016$.

The results in Fig. 3 show that all curves displays similar behavior, meaning that the precision increases as more future edges are evaluated. Therefore, if we consider larger time scale, the tendency is that the most similar authors will indeed be linked by a citation link. The behavior is also independent of the considered test set and similarity measurements: highly similar edges are predicted with high precision when considering larger periods. The precision slowly drops as similar authors are not linked by citations when considering shorter future time windows.

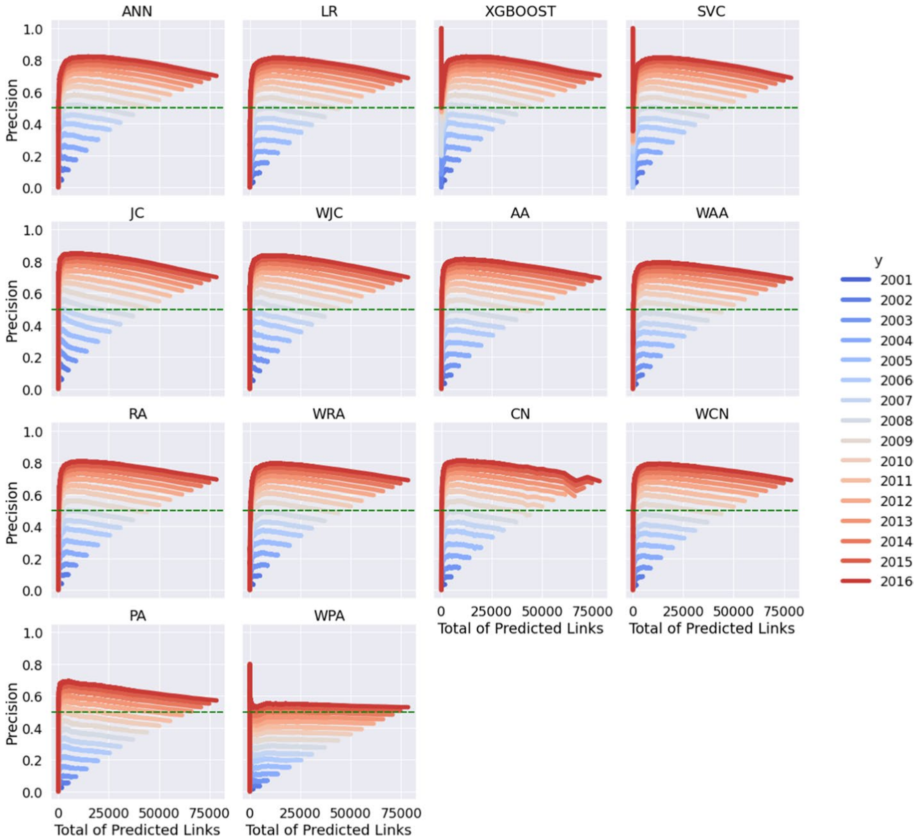

Fig. 3 Evolution of precision values as the most similar edges are included in the link prediction task. In each subpanel, each curve corresponds to a different validation set. Curves ranging from blue to red correspond to validation sets considering papers published in the interval [2000, y]. Green dashed lines correspond to the expected precision when links are randomly placed

One interesting finding from Fig. 3 is that both preferential attachment similarity measurements are clearly outperformed by the other measurements. This means that neighborhood information plays an important role in the task of predicting citations in author citation networks. This lack of performance confirms that the prediction of both edge ends in an author citation network is not trivially performed with the preferential attachment rule. We should note, though, that the PA rule is a strong predictor of how many citations a researcher will accrue in the future (Silva et al. 2020). The number of future citations depends mostly on the number of citations received in the last 12-24 months (Silva et al. 2020). The reasons for not citing similar structural authors are two-fold: authors have a limited vision of the network structure, which may cause them to miss the papers of other authors. While semantical dissimilarity could be a different reason for similar structural authors not citing each other, it should be mentioned that even highly semantical similar papers are frequently overlooked when authors perform a systematic review (Amancio et al. 2012a).

The precision for selected quantities of included edges is also shown in Table 1. The results obtained for the Jaccard Index (both unweighted and weighted versions) confirmed that this measurement is the most effective measurement to predict future citations, achieving 84.9% and 83.5% of precision when predicting the top 10,000 top edges, respectively for JC and WJC. Both metrics were significantly better than the other considered approaches. Surprisingly, the approach based on artificial neural networks (ANN) and XGBoost were outperformed by the simple Jaccard index. When predicting the largest amount of link, however, all methods displayed similar performance. Both Adamic-Adar and Common Neighbors metrics also yield good results, especially when predicting a larger number of edges. The weighted version of Adamic-Adar, Common Neighbors and

Table 1 Comparison of precision values when considering all individual similarity measurements and the learning models in the 2000–2016 validation dataset

Method	Precision 2×10^3	Precision 10×10^3	Precision 20×10^3	Precision 30×10^3	Precision 60×10^3	Precision 78, 569
JC	**0.831**	**0.849**	**0.839**	**0.824**	**0.750**	0.700
WJC	0.804	0.835	0.832	0.816	0.749	0.700
ANN	0.752	0.818	0.821	0.808	0.748	0.701
XGBOOST	0.754	0.818	0.820	0.810	0.742	**0.702**
AA	0.778	0.813	0.808	0.795	0.738	0.695
CN	0.786	0.814	0.805	0.791	0.736	0.684
SVC	0.759	0.812	0.812	0.799	0.74	0.689
LR	0.756	0.811	0.810	0.798	0.738	0.687
RA	0.760	0.807	0.803	0.791	0.732	0.694
WAA	0.741	0.791	0.789	0.778	0.729	0.691
WCN	0.743	0.791	0.787	0.776	0.729	0.690
WRA	0.733	0.794	0.789	0.779	0.728	0.690
PA	0.684	0.678	0.666	0.648	0.598	0.572
WPA	0.534	0.551	0.549	0.546	0.535	0.529

The methods are ordered in descending order, from top to bottom, by the precision average. Each column represents the precision obtained when different number of edges were included in the link prediction analysis. The total number of included edges varied between 2000 to 78, 569 edges, which represents to total of authors citations between 2000 and 2016. The best precision is highlighted for each column

Resource Allocation, compared to the other methods displayed a lower performance. These results reinforce the fact that, when precision is sought, different measurements based on the same information (i.e. common neighbors) can lead to distinct performance. Finally, one can observe that a low precision was observed for the preferential attachment method. Even when predicting the most similar 2, 000 edges, the performance is comparable to a random classification. The weighted version performs even worse.

In sum, there is no significant improvement in performance when using such measurements as a summarization measurement, when compared with the performance obtained via Jaccard index. This is consistent with recent literature showing that neural networks can also be outperformed by traditional classifiers (Amancio et al. 2014). Even though graph neural networks have become popular in recent years, they require a broader knowledge of the network structure (Cui et al. 2018), which implies a much more expensive computation time.

True vs. False positive analysis

While in the previous section we focused on precision, here we compare the methods by considering the *receiver operating characteristic* (ROC) curve (Davis and Goadrich 2006). The ROC curve establishes a relationship between true positive and false positive rates. Thus, higher values of AUC are expected whenever true positives are more frequently identified than false positives (Davis and Goadrich 2006) as the threshold in similarity for including new links decreases. This analysis is important because, differently from the precision, the AUC curve also considers the efficiency of the model in *not predicting* links that *will not exist*.

The ROC curves obtained for each of the considered measurements are illustrated in Figure S1 of the Supporting Information. The corresponding AUC curves are shown in Fig. 4. Differently from the precision analysis, a higher performance is observed when predicting links established within short and long periods. In other words, the efficiency drops when predicting both the existence and absence of links in a mid-term scale. Interestingly, when considering the intermediate test dataset [2000 – 2007], apart from the PA measurement, all measurements have similar performance. The highest differences arise when predicting citations established in long-term periods. In this context, we do observe a

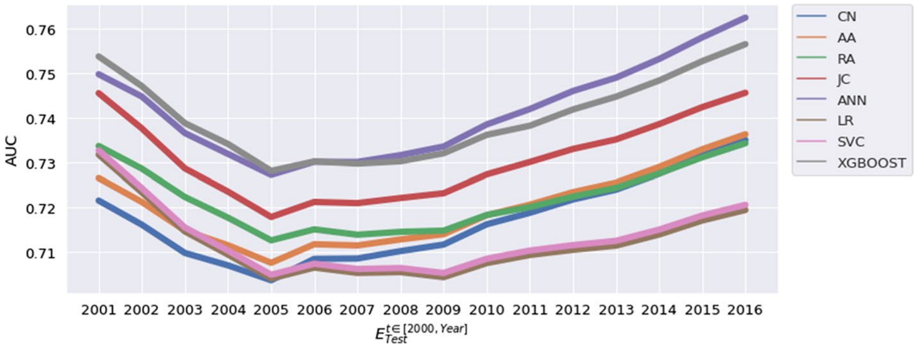

Fig. 4 *Area Under the ROC Curve* (AUC) for all the similarity measurements considered in our analysis by each year used in the test dataset. Overall, the best performance was obtained with the artificial neural networks (ANN) and XGBoost

difference in performance. When predicting citations that were established within the first year (2001), the best performances were achieved with ANN and XGBoost. The Jaccard measurement was found to be the third the most accurate metric, regardless of the end year. Interestingly this result also reinforces the effectiveness of the Jaccard Index in a wider context, since this same measurement has been reported to be relevant in predicting *paper citations*, even when compared to other global measurements (Shibata et al. 2012).

Some insights can be drawn from the observed results. Once again, preferential attachment-based measurements turned out to be the ones yielding the lowest performance (result not shown in Fig. 4). Because PA and WPA do not consider the distance between nodes to predict links, this result might indicate that, for most of the future citations at the author level, the local similarity seems to play an important role to gauge nodes similarity. In other words, while in many real-world we do know that there is a higher chance of two hubs to be connected, this type of information alone is not very useful to predict citations in author citation networks. The semantic and geographic component are important features that are being disregarded when only nodes connectivity is considered (Hennemann et al. 2012; Amancio et al. 2012b). More informed measurements using the degree of shared neighbors to quantity similarity includes AA, RA and their weighted versions. These measurements take high values whenever there are several shared neighbors and those neighbors do not share many links with other nodes other than the ones being evaluated. Because AA and RA displayed low performance, the degree of neighbors, therefore, also do not bring much information regarding similarity. It remains to be shown, however, if the degree of intermediary nodes connecting the authors being evaluated is also not useful when *longer connection paths* are considered.

Another interesting result is that the weighted versions of the considered measurements were not able to significantly improve the performance of the link prediction task (result not shown).

In Table 2 we summarize the results obtained in Fig. 4. We show the mean, minimum, maximum and quartile values obtained for each measurement by considering the distribution of AUC values along the considered years. ANN and XGBoost yielded similar values in all considered metrics, being the methods achieving highest performance, as observed in Fig. 4. The results confirm that the combination of network similarity metrics via neural network and XGBoost provides results that are better than the one obtained when using a single similarity metric. Such an optimized performance is not evident, however, when the metrics are combined via other classifiers.

Relationship between unweighted and weighted metrics

Here we analyzed the correlation between of the considered measurements. This analysis is important because highly correlated measurements might share the same information and thus, their combination can lead to a minimum gain in performance.

We found in the previous sections that the weighted version of the considered measurements does not significantly improve the performance of the link prediction task. The reason for this could be the fact that the weight information is not relevant for the task. A similar performance could stem from the possibility of the weighted and unweighted version of the same metric being correlated. In this section, we probe whether this hypothesis could be playing a role in the citation prediction task.

In Fig. 5 we show the Pearson correlation values between the unweighted and the respective weighted version of the traditional local network similarity indexes. Apart from

Table 2 Comparison of Area Under the Curve (AUC) values when considering all the similarity measurements and the learning models

Method	AUC Mean	AUC Min	AUC Q1	AUC Q2	AUC Q3	AUC Max
ANN	**0.742**	0.727	**0.732**	**0.740**	**0.749**	**0.762**
XGBOOST	0.740	**0.728**	0.732	0.738	0.747	0.756
JC	0.731	0.718	0.723	0.729	0.738	0.746
WJC	0.730	0.718	0.721	0.729	0.737	0.748
RA	0.722	0.713	0.715	0.721	0.728	0.734
AA	0.720	0.708	0.713	0.719	0.726	0.736
WRA	0.719	0.711	0.712	0.717	0.724	0.735
CN	0.717	0.704	0.710	0.716	0.722	0.735
WAA	0.716	0.706	0.709	0.715	0.722	0.731
WCN	0.715	0.704	0.708	0.713	0.720	0.730
SVC	0.713	0.705	0.707	0.711	0.716	0.733
LR	0.712	0.704	0.706	0.710	0.715	0.732
PA	0.590	0.584	0.586	0.592	0.593	0.597
WPA	0.549	0.538	0.541	0.550	0.556	0.559

Each column represents a summary of the results obtained along the years. Rows are ordered by AUC mean. To summarize the AUC we used the mean, minimum, quartiles (Q1–Q3) and the maximum value. The best precision is highlighted for each column

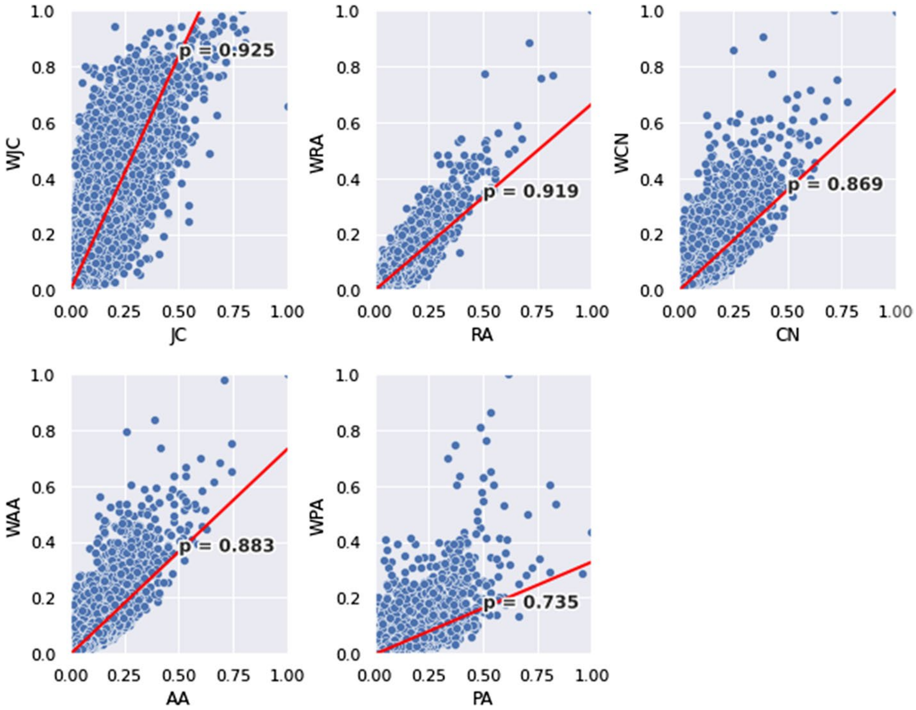

Fig. 5 Correlation analysis between the unweighted and weighted versions of the local network similarity metrics

the the PA metric, we found in all cases high correlation values. The highest correlation was found with the Jaccard metric, which is the method yielding high performance in the task. Therefore, for the link prediction task in author citation networks, edges weight does not seem to yield additional information.

While the correlation between the unweighted and weighted version of the PA metric is still high, the value is lower than the correlation found for the other metrics. This means that the weight, in this case, provides additional information that can not be recovered via the number of links alone. However, despite this difference, the PA metric is significantly outperformed by other metrics in the citation prediction task. In other words, even though additional information can be found when including edges weight, the provided information is irrelevant for the task.

Conclusion

While link prediction have been widely studied in scientometrics scenarios, the analysis and comparisons of methods have been mostly limited to predicting links in paper citation networks and other limited scenarios at the author level (Daud et al. 2017). Here we performed a systematic performance comparison of several local similarity measurements in the context of predicting links in *authors citation networks*. Because name ambiguity is a major problem when dealing with names in scientometrics datasets, we used a disambiguated dataset of names provided by the *Microsoft Academic Graph* (Wang et al. 2020).

Our comparative analysis focused on local network information to avoid the complexity of analyzing very large networks. While local measurements indeed may not achieve state of the art performance, given the limited information they rely on, such techniques have shown to yield good performance while not being computationally costly (Shibata et al. 2012). In addition to the traditional network similarity measurements, we also used extensions of these measurements that consider the weighted nature of author citation networks. Our comparative approach revealed several interesting results. The Jaccard Index turned out to be the most effective similarity index, when compared with other traditional network similarity measurements. We found that the preferential attachment rule alone is not informative for the task, despite the fact that the total number of citations received by authors is well described by preferential attachment rules (Silva et al. 2020). Surprisingly, the considered Logistic Regression and Support Vector Machine Classifier techniques did not yield the best results. This suggests that these machine learning strategy, when used to summarize information extracted by local network similarity measurements, is not competitive to predict future citations between authors. Besides that, both Artificial Neural Networks and XGBoost displayed the best performance. Our analysis also revealed that, apart from the Preferential Attachment rule, the use of edges weight does not significantly improves the performance of the task because the weighted and unweighted version of the metrics are strongly correlated.

Predicting citations is a task that encompasses many different factors, and thus the accuracy of the model depends both on which information is used to address the problem and the technique employed to find patterns in the data. Even at the paper level, many factors may affect citations, which makes it hard to predict even the number of citations a paper will accrue (Bai et al. 2019). Here we focused on a simple information regarding authors citations networks. We take the view that the best performance – 84.9% of precision—is not a weak result given the simplicity of the used similarity metrics. We expect, however, that the use of additional information, such as semantical similarity, global network similarity metrics and other metadata can improve the performance of the task. However, it is still important to have in mind that other type of information might be already encoded in citation links. This is the case e.g. of the textual information, since a fraction of papers linked by citations are semantically similar (Amancio et al. 2012a).

While in this paper we focused on predicting authors' citations, we stress that counting citations should not be the only metric used to evaluate researchers. As a side effect, predicting links with high precision in author citation networks could be useful to detect the authors that will receive more citations in the future. This does not mean, however, that one would be able to predict the relevance of authors' research output. Many studies have pointed out that citations should not be treated equally, and thus the number of citations should not be used as a unique metric to measure scientific quality (Parnas 2007; Edwards and Roy 2017). Several works advocate that one should use the citation context to better understand if a citation link can be related to research quality and impact (Zhang et al. 2013; Bornmann and Daniel 2008).

This paper focused only on local information to predict new links, since a local analysis mitigates the cost of computing pairwise similarity indexes via global information. In future works, other extensions could be considered in a comparative analysis. One could introduce further hierarchies when comparing neighbors. However, this additional complexity could also lead to noise since many higher-level neighbors might be shared by many authors due to the small-world effect (Hung and Wang 2010). As a consequence, the characterization performance might decrease with the introduction of deeper concentric circles (Amancio et al. 2011). This could be addressed, by providing a lower relevance to higher hierarchies by using a strategy (see e.g. Amancio et al. (2015)) that linearly combines similarity values observed in both first and higher hierarchical levels. In a similar direction, information from further hierarchies could be introduced via recent network embeddings techniques, where nodes can be artlessly compared via vector similarity measurements.

Regarding the limitations and potential future research that could spark from the current study, one could extend the number of authors considered. We focused our analysis on the most prolific authors, since they are more active in the field and thus the training and evaluation do not suffer from the lack of data. The expansion of the number of authors in the considered dataset could lead to potential improved results, since more information is provided to the methods. Another possibility to improve the results is to capture citations from external datasets. While enriching the network can be useful to unveil hidden similarities in the current dataset, this could however lead to spurious information if authors' name are not matched in an accurate way. Another possible extension consists in analyzing whether different subfields are more predictable in the sense that citations can be predicted with higher accuracy. The results could be potentially relevant in the conception of different strategies to measure similarity in different subfields of science.

In addition to using additional sources of citation data, the usage of external data to stack on top of the information already provided by the author citation network could potentially improve the machine learning results. For example, one could taking into account the geographical distance between authors, since many collaborations occur within short distances (Hennemann et al. 2012). The research topic is also a factor that could be incorporated in the model, since one should expect that scientists sharing interest in the same research topic are more likely to have a citation link connecting them. At the semantic level, bag-of-words, complex networks and/or neural text representations could be used to calculate the similarities between authors (Amancio et al. 2012b; Katz 1994; Wuestman et al. 2019; Stella 2019, 2020). All different sources of information could be combined e.g. using a linear combination to generate a network reflecting all the different attributes (Amancio et al. 2015).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-022-04484-6.

Acknowledgements A preprint version of this manuscript is available at arXiv (Vital and Amancio 2021). D.R.A. acknowledges financial support from São Paulo Research Foundation (FAPESP Grant No. 2020/06271-0) and CNPq-Brazil (Grant No. 304026/2018-2 and 311074/2021-9). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Declarations

Conflict of interest The authors have no conflict of interest to declare.

References

Adamic, E., & Adar, LA. (2003). Friends and neighbors on the web (3):211–230

Amancio, D. R., Nunes, M. G. V., Oliveira, O. N., Jr., Pardo, T. A. S., Antiqueira, L., & Costa, L. F. (2011). Using metrics from complex networks to evaluate machine translation. *Physica A: Statistical Mechanics and its Applications, 390*(1), 131–142.

Amancio, D. R., Nunes, Md. G. V., Oliveira, O. N., Jr., & da F Costa L,. (2012). Using complex networks concepts to assess approaches for citations in scientific papers. *Scientometrics, 91*(3), 827–842.

Amancio, D. R., Oliveira, O. N., Jr., & da Fontoura, Costa L. (2012). Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index. *Journal of Informetrics, 6*(3), 427–434.

Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & Costa, L. F. (2014). A systematic comparison of supervised classifiers. *PLoS One, 9*(4), e94. 137.

Amancio, D. R., Oliveira, O. N., Jr., & Costa, Ld. F. (2015). Topological-collaborative approach for disambiguating authors' names in collaborative networks. *Scientometrics, 102*(1), 465–485.

Bai, X., Xia, F., Lee, I., Zhang, J., & Ning, Z. (2016). Identifying anomalous citations for objective evaluation of scholarly article impact. *PloS One, 11*(9), e0162.

Bai, X., Zhang, F., & Lee, I. (2019). Predicting the citations of scholarly paper. *Journal of Informetrics, 13*(1), 407–418.

Bai, X., Zhang, F., Ni, J., Shi, L., & Lee, I. (2020). Measure the impact of institution and paper via institution-citation network. *IEEE Access, 8*, 548–555.

Barabási, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications, 311*(3–4), 590–614. https://doi.org/10.1016/s0378-4371(02)00736-7

Bornmann, L., & Daniel, HD. (2008). What do citation counts measure? a review of studies on citing behavior. Journal of documentation

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Chacon, X. S., Silva, T. C., & Amancio, D. R. (2020). Comparing the impact of subfields in scientific journals. *Scientometrics, 125*(1), 625–639.

Chen, S., Dang, D., Macy, R., & Rockwell, C. (2019). Link prediction on the patent citation network. https://crockwell.github.io/data/LP_patent.pdf

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp 785–794

Cui, P., Wang, X., Pei, J., & Zhu, W. (2018). A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering, 31*(5), 833–852.

Daud, A., Ahmed, W., Amjad, T., Nasir, JA., Aljohani, NR., Abbasi, RA., & Ahmad, I. (2017). Who will cite you back? reciprocal link prediction in citation networks. Library Hi Tech

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In: Proceedings of the 23rd international conference on Machine learning, pp 233–240

Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science, 34*(1), 51–61.

Eom, Y. H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLoS ONE, 6*(9), e24-926.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., & Barabási, A. L. (2018). Science of science. *Science*. https://doi.org/10.1126/science.aao0185

Hennemann, S., Rybski, D., & Liefner, I. (2012). The myth of global science collaboration-collaboration patterns in epistemic communities. *Journal of Informetrics, 6*(2), 217–225.

Hug, SE., & Brändle, MP. (2017). The coverage of microsoft academic: Analyzing the publication output of a university. CoRR arxiv:bs/1703.05539

Hung, S. W., & Wang, A. P. (2010). Examining the small world phenomenon in the patent citation network: a case study of the radio frequency identification (rfid) network. *Scientometrics, 82*(1), 121–134.

Jain, A., Mao, J., & Mohiuddin, K. (1996). Artificial neural networks: a tutorial. *Computer, 29*(3), 31–44. https://doi.org/10.1109/2.485891

Katz, J. (1994). Geographical proximity and scientific collaboration. *Scientometrics, 31*(1), 31–43.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), 1995*, 1137–1145.

Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., & Hütt, M. T. (2011). Motifs in co-authorship networks and their relation to the impact of scientific publications. *The European Physical Journal B, 84*(4), 535–540.

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 pp 1–19

Li, W., Aste, T., Caccioli, F., & Livan, G. (2019). Reciprocity and impact in academic careers. *EPJ Data Science, 8*(1), 20.

Liu, X. F., Chen, H. J., & Sun, W. J. (2021). Adaptive topological coevolution of interdependent networks: Scientific collaboration-citation networks as an example. *Physica A: Statistical Mechanics and its Applications, 564*(125), 518.

Lü, L., & Zhou, T. (2010). Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters), 89*(18), 001.

Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications, 390*(6), 1150–1170.

Martinčić-Ipšić, S., Močibob, E., & Perc, M. (2017). Link prediction on twitter. *PLoS ONE, 12*(7), 1–21. https://doi.org/10.1371/journal.pone.0181079

Milojević, S. (2013). Accuracy of simple, initials-based methods for author name disambiguation. *Journal of Informetrics, 7*(4), 767–773.

Molléri, J. S., Petersen, K., & Mendes, E. (2018). Towards understanding the relation between citations and research quality in software engineering studies. *Scientometrics, 117*(3), 1453–1478.

Newman, M. (2001). Clustering and preferential attachment in growing networks. *Physical Review E, 64*(2), 025–102.

Nie, Z., Liu, Y., Yang, L., Li, S., & Pan, F. (2021). Construction and application of materials knowledge graph based on author disambiguation: Revisiting the evolution of lifepo4. Advanced Energy Materials p 2003580

Nielsen, M. A. (2015). *Neural networks and deep learning* (Vol. 25). CA: Determination press San Francisco.

Nielsen, M. W., & Andersen, J. P. (2021). Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences, 118*(7), 2012208118.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology, 24*(12), 1565–1567.

Parnas, D. L. (2007). Stop the numbers game. *Communications of the ACM, 50*(11), 19–21.

Powell, W. W., White, D. R., Koput, K. W., & Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology, 110*(4), 1132–1205.

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E, 80*(5), 056–103.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of Database Systems, 5*, 532–538.

de Sá, H., & Prudencio, R. (2011). Supervised link prediction in weighted networks. In: Neural Networks (IJCNN), The 2011 International Joint Conference on, IEEE, pp 2281–2288

Sebo, P., de Lucia, S., & Vernaz, N. (2021). Accuracy of pubmed-based author lists of publications and use of author identifiers to address author name ambiguity: a cross-sectional study. Scientometrics pp 1–15

Shibata, N., Kajikawa, Y., & Sakata, I. (2012). Link prediction in citation networks. *Journal of the American Society for Information Science and Technology, 63*(1), 78–85.

Silva, F. N., Amancio, D. R., Bardosova, M., Costa, Ld. F., & Oliveira, O. N., Jr. (2016). Using network science and text analytics to produce surveys in a scientific topic. *Journal of Informetrics, 10*(2), 487–502.

Silva, F. N., Tandon, A., Amancio, D. R., Flammini, A., Menczer, F., Milojević, S., & Fortunato, S. (2020). Recency predicts bursts in the evolution of author citations. *Quantitative Science Studies, 1*(3), 1298–1308.

Stella, M. (2019). Modelling early word acquisition through multiplex lexical networks and machine learning. *Big Data and Cognitive Computing, 3*(1), 10.

Stella, M. (2020). Multiplex networks quantify robustness of the mental lexicon to catastrophic concept failures, aphasic degradation and ageing. *Physica A: Statistical Mechanics and Its Applications, 554*(124), 382.

Vital, Jr A., & Amancio, DR. (2021). A comparative analysis of local network similarity measurements: application to author citation networks. arXiv:2103.13946

Wang, K., Shen, Z., Huang, C., Wu, C. H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies, 1*(1), 396–413.

Wang, M., Yu, G., & Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications, 387*(18), 4692–4698.

Wang, P., Xu, B., Wu, Y., & Zhou, X. (2014). Link prediction in social networks: the state-of-the-art

Wright, RE. (1995). Logistic regression.

Wuestman, M. L., Hoekman, J., & Frenken, K. (2019). The geography of scientific citations. *Research Policy, 48*(7), 1771–1780.

Yegnanarayana, B. (2009). *Artificial neural networks*. Delhi: PHI Learning Pvt. Ltd.

Zhang, G., Ding, Y., & Milojević, S. (2013). Citation content analysis (cca): A framework for syntactic and semantic analysis of citation content. *Journal of the American Society for Information Science and Technology, 64*(7), 1490–1503.

Zhang, L., & Ban, Z. (2020). Author name disambiguation based on rule and graph model. In: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, pp 617–628

Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. *The European Physical Journal B, 71*(4), 623–630.