EPJ.org

O EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE   **Open Access**

Prediction of new scientific collaborations through multiplex networks

Marta Tuninetti[1][†], Alberto Aleta[1][†], Daniela Paolotti[1], Yamir Moreno[1,2,3] and Michele Starnini[1][*] (iD)

[*]Correspondence:
michele.starnini@gmail.com
[1]ISI Foundation, via Chisola 5, 10126
Torino, Italy
Full list of author information is
available at the end of the article
[†]Equal contributors

Abstract

The establishment of new collaborations among scientists fertilizes the scientific environment, fostering novel discoveries. Understanding the dynamics driving the development of scientific collaborations is thus crucial to characterize the structure and evolution of science. In this work, we leverage the information included in publication records and reconstruct a categorical multiplex networks to improve the prediction of new scientific collaborations. Specifically, we merge different bibliographic sources to quantify the prediction potential of scientific credit, represented by citations, and common interests, measured by the usage of common keywords. We compare several link prediction algorithms based on different dyadic and triadic interactions among scientists, including a recently proposed metric that fully exploits the multiplex representation of scientific networks. Our work paves the way for a deeper understanding of the dynamics driving scientific collaborations, and validates a new algorithm that can be readily applied to link prediction in systems represented as multiplex networks.

Keywords: Scientific collaboration networks; Computational social science; Link prediction

1 Introduction

One of the main drivers of scientific discoveries is the establishment of new collaborations among researchers. The collision of different scientific trajectories, even if they belong to the same research area, brings together different methods, concepts, and ideas, fostering the ideal environment for scientific creativity. Understanding the dynamics that drives the development of scientific collaborations is thus pivotal to characterize the structure and evolution of science [1]. In this endeavour, two factors play a crucial role. On the one hand, the digitalization of large scale bibliographic databases has provided comprehensive data sets of publication records including research from all disciplines, without geographical limits. By leveraging on these databases [2], researchers have pictured the structure of different research fields [3, 4], measured the emergence of new interdisciplinary areas [5, 6], mapped the evolution of scientific interests [7, 8], and characterized scientific productivity at the individual and geographical level [9–12]. On the other hand, network science [13] has been established as the main tool to analyze and model cooperation in science. Since the seminal work by Newman [14], scientific collaborations are represented in the form of

Springer

a network, where nodes stand for scientists and a link between two nodes is drawn if two scientists have co-authored a paper together.
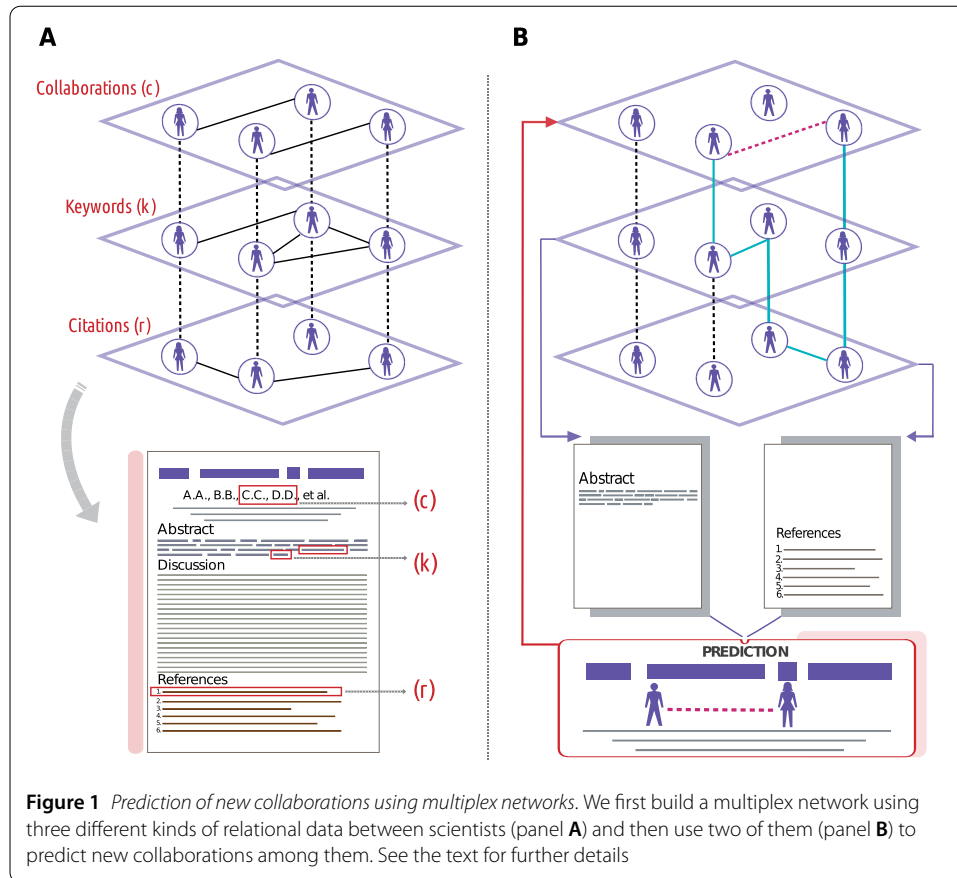
Forecasting a new collaboration translates, within the network science domain, into a link prediction problem [15], a prolific area of network research with applications ranging from detecting hidden links in economic networks [16] to enhancing user experience in online social platforms [17]. Many link prediction algorithms are based on similarity measures computed on the node attributes, i.e. two nodes are likely to be linked if they are similar with respect to certain features [18]. In social networks, one of the most successful similarity metrics for link prediction is the presence of common neighbors between two nodes, e.g. a new friendship on Facebook can be recommended on the basis of the number of common friends shared [19, 20]. Despite its simplicity, this concept has proven to be quite successful for link prediction in scientific collaboration networks [21, 22]. Moving from this simple approach, several attempts have been made to improve the prediction of collaborations by incorporating additional layers of data, for instance, by adding information about the organization the authors work at [23], topical interest [24], time at which collaborations are established [25], offline relationships among employees of the same university [26], weights of the collaboration links [27] or journal information [28]. However, in most of these approaches, the scores are computed individually for each set of data and then aggregated into a unique score, possibly after associating a specific weight to each set.

In this paper, we merge different bibliographic sources to leverage the whole information included in publication records to improve the prediction of new collaborations. To this aim, we reconstruct a multiplex network [29, 30] in which nodes represent scientists and different kinds of relations among them are encoded in different layers, i.e., a given relational category corresponds to a layer, see Fig. 1(A). In particular, we focus on scientific credit, represented by citations, and common interests, measured by the usage of common keywords, to predict new collaborations. We compare several link prediction algorithms based on different dyadic and triadic interactions between scientists. We also consider a recently proposed metric for link prediction in multiplex networks, based on a generalization of the Adamic-Adar method for single-layered networks [31], able to fully exploit the multiplex representation of scientific networks. We show that scientific credit and common scientific interests can be predictive of new collaborations between scientists.

2 Data

Our dataset is composed by merging two different bibliographical sources. First, the American Physical Society (APS) database, including authors' names, publication date and references of over 400,000 papers published from 1893 to 2009 [32]. Here we considered the disambiguated dataset published in Ref. [12]. Second, the ArnetMiner database [33], containing title, authors' list, publication year and keywords for almost 155 billion papers belonging to multiple research fields. From this dataset, we select only those papers present in the APS dataset, by matching the DOI number. Our final dataset is composed, for each paper, by the list of authors with their affiliations, the list of keywords associated to the paper, and the papers cited as references. Before analyzing the data, we apply a cleaning procedure to the information related to keywords, see Additional file 1 for details.

We then reconstruct a scientific weighted multiplex network [34], where nodes represent scientists and different layers account for different interactions among them: collaborations, common interests, and scientific credit, see Fig. 1(A). A first layer (*c*) represents

Figure 1 *Prediction of new collaborations using multiplex networks.* We first build a multiplex network using three different kinds of relational data between scientists (panel **A**) and then use two of them (panel **B**) to predict new collaborations among them. See the text for further details

collaborations and it corresponds to a classical co-authorship network: two authors are linked if they published at least a paper together. A second layer (r) represents scientific credit, measured by references or citations: a link from author u to author v indicates that u cited at least one paper from v. Lastly, the third layer (k) represents common scientific interests, which can be measured by the usage of common keywords: two authors are connected if, out of all the keywords they have ever used, they have at least one in common. The collaboration and keyword layers are formed by undirected links, while the reference layer includes directed interactions. Finally, the weight w_{uv}^{α} of a link represents the number of co-authored papers ($\alpha = c$), citations ($\alpha = r$), or common keywords ($\alpha = k$) between two authors u and v.

We consider two subsequent time intervals, first an interval over which link prediction algorithms will be trained, corresponding to a training network with all authors who published a paper between t_0 and t_1, and a test interval for testing the predictions of new collaborations, including all authors active between t_1 and t_2. We then consider the prediction of new links in a subset of nodes of these networks, which we name *Core*, corresponding to the authors that have at least $k_{\min}$ edges in the collaboration layer, i.e., a minimum number of co-authors equal to $k_{\min}$ both in the training and test intervals. This choice is to ensure authors to be active in both intervals, as it is common practice in link prediction problems on social networks [15]. In order to reduce the computational complexity of the prediction algorithms, we restrict our analysis to only papers published in Physical Review Letters (PRL) between $t_0 = 1994$ and $t_2 = 2005$, split at $t_1 = 2000$, see Ad-

Table 1 Properties of the different layers of the scientific networks and the *Core* over which link prediction is computed. We show the number of nodes N, the total weight $W = \sum_{ij\alpha} w_{ij}^{\alpha}$, the average degree $\langle k \rangle$, the overlap between the collaboration layer and the other layers, and the global clustering coefficient C. The overlap is defined as the fraction of links in the collaboration layer that are also present in citations layer or keyword layers

Network	N	W	$\langle k \rangle$	Overlap	C
Collaborations layer	24,366	78,082	5.42	–	0.5
Citations layer	24,366	275,260	8.9	0.33	0.24
Keywords layer	24,366	28,872,677	1998	1	0.44
Core	5944	17,635	4.27	–	0.45

ditional file 1 for details on how we choose the intervals. The resulting scientific multiplex network thus includes only authors, citations, and keywords related to papers published in PRL and is composed by N = 24,366 authors. By setting $k_{min} = 3$, the *Core* set for link prediction is composed by 5944 nodes, while the number of new links to be predicted is equal to $E_p = 7563$. In Additional file 1, we show results for link prediction with a *Core* obtained by setting $k_{min} = 5$. Table 1 reports several properties of the different layers of the scientific multiplex network and the *Core* over which link prediction is computed. In particular, note that the keyword layer is denser than the others.

3 Link prediction algorithms

To determine if the information provided by the citation and keyword layers is actually useful to predict the appearance of new links in the collaboration layer (see Fig. 1B), we propose several novel metrics based on the similarity between nodes in these layers.

First, we consider metrics based on dyadic interactions between scientists, that is, to predict a new collaboration between nodes u and v (i.e., a new $u - v$ link in the collaboration layer), we consider links between nodes u and v in different layers. For instance, we consider Mutual Citations (*MC*): if two authors mutually cite each other, it might be more likely for them to collaborate. The *MC* score between nodes u and v is defined simply as the weight of the link between u and v in the citation layer,

$$MC(u,v) = w_{uv}^r. \tag{1}$$

Similarly, we consider Common Keywords (*CK*): if two authors show common scientific interests, using the same set of keywords, the chances that they collaborate in the future should be higher than if they did not have common interests. Thus, the *CK* score between nodes u and v can be expressed as the weight of a link between u and v in the keyword layer,

$$CK(u,v) = w_{uv}^k. \tag{2}$$

For each case, we also define a normalized variant. The Normalized Mutual Citations (*NMC*) score normalizes the number of citations between two authors by the total citations received by each of them. The idea is that mutual citations between very popular scientists (who attract many citations in general) should count less than mutual citations between scientists receiving less citations. The *NMC* is thus defined as

$$NMC(u,v) = \frac{w_{uv}^r}{s_u^r} + \frac{w_{vu}^r}{s_v^r}, \tag{3}$$

where $s_u^r = \sum_v w_{vu}^r$ is the total number of citations received by u, corresponding to the total incoming strength. Note that this metric considers the directed citation network, explicitly differentiating between incoming and outgoing citations. The last dyadic metric considered is the Normalized Common Keywords (NCK), computed as

$$NCK(u,v) = \frac{w_{uv}^k}{\max(K_u, K_v)}, \tag{4}$$

where K_u is the keyword list used by node u. Here, the idea is that authors using more keywords than others are more likely to share keywords with someone else.

Next, we consider metrics based on triadic closure. That is, to predict a new $u - v$ link in the collaboration layer, we consider triangles involving nodes u and v in different layers. The most common and successful method of this class has been developed by Adamic and Adar (AA) [19]. The AA score between nodes u and v is given by counting their common neighbors w weighted by the inverse of the logarithm of their degree. In this way, more active authors, which are more likely to be common neighbors of a given pair of nodes, weight less in the AA score. In a multiplex network, the AA score could be applied to different layers, i.e. considering neighbors also in layers different from the one where new links are predicted. Therefore, the AA score computed by counting neighbors in layer α can be defined as

$$AA_\alpha(u,v) = \sum_{w \in \Gamma_\alpha(u) \cap \Gamma_\alpha(v)} \frac{1}{\ln(k_w^\alpha)}, \tag{5}$$

where $\Gamma_\alpha(u)$ represents the set of neighbors of node u in layer α and $k_w^\alpha = |\Gamma_\alpha(w)|$ is the degree of node w in layer α. By applying Eq. (5) to the collaboration layer ($\alpha = c$), one has the classical AA score for collaboration networks, AA_c: two scientists are more likely to collaborate if they share many common collaborators. Equation (5) can also be applied to the citation ($\alpha = r$) or keyword ($\alpha = k$) layer, the rationale being that two scientists are more likely to collaborate if they cite the same set of authors (AA_r score for the citation layer), or have similar scientific interests (AA_k score for the keyword layer). Note that in all cases, the common neighbors of u and v can be both in the *Core* and outside it.

Finally, we consider a recently proposed generalization of the AA score [19] to multiplex networks, which takes into account all possible triadic closures in multiplex networks [31]. The MAA score for the prediction of a link between nodes u and v in the collaboration layer c is defined as

$$MAA(u,v) = \sum_{\alpha,\beta} \sum_{w \in \mathcal{T}_{\alpha\beta}} \frac{\eta_{c\alpha} \eta_{c\beta}}{\sqrt{\langle k \rangle_\alpha \langle k \rangle_\beta}} \frac{1}{\sqrt{\ln(k_w^\alpha) \ln(k_w^\beta)}}, \tag{6}$$

where $\mathcal{T}_{\alpha\beta}$ are different kinds of triadic relations among three nodes u, v and w [31]. While the link $u - v$ to be predicted is in the collaboration layer, the other two links $u - w$ and $v - w$ may lay in any layer. For instance, one link $u - w$ in the collaboration layer ($\alpha = c$) and the other link $v - w$ in the citation ($\beta = r$) or keyword ($\beta = k$) layer, or one link $u - w$ in the citation layer ($\alpha = r$) and the other link $v - w$ in the keyword ($\beta = k$) layer. The coefficients $\eta_{c\alpha}$ and $\eta_{c\beta}$ before each term control the relative weight of each type of triadic closure in the total score of the link, thus $\eta_{c\alpha}$ corresponds to the weight of layer α, with

$\sum_\alpha \eta_{c\alpha} = \eta_{cc} + \eta_{cr} + \eta_{ck} = 1$. The case $\eta_{cc} = 1$, $\eta_{cr} = \eta_{ck} = 0$, corresponds to the classical AA_c score on collaboration networks, while $\eta_{cr} = 1$ ($\eta_{ck} = 1$) corresponds to the AA_r (AA_k) score applied to the citation (keyword) layer, see Fig. 1 for a schematic illustration of this process.

4 Results

The quality of link prediction algorithms is usually evaluated by the Receiver Operating Characteristics (ROC) curve, with the corresponding Area Under the Curve (AUC) value. However, due to the limited amount of links present in a network, the AUC of any similarity-based link prediction algorithm is bounded [31, 35]. For this reason, we also consider the Precision of different scores, computed as n^*/n, where n is the number of new links that we want to predict and n^* is the amount of correct predictions among the top n links.

As a first step, we explore the coefficients $\eta_{c\alpha}$ of the *MAA* metric to find the combination that maximizes the prediction of new collaborations. Figure 2 shows the AUC and Precision of the *MAA* metric, given by Equation (6), as a function of the coefficients $\eta_{c\alpha}$. Figure 2(a) shows that the AUC value has an important contribution from triads involving the citations and keywords layers, as shown by the discontinuity for $\eta_{cc} < 1$. This result is consistent with the fact that citations and keywords relationships contribute to increase the amount of information carried by the collaboration layer, see Table 2. The Precision is maximum for $\eta_{ck} = 0.05$ and $\eta_{cr} = 0.1$ (see Fig. 2(b)), showing that the contribution of the collaboration layer is important to keep high precision. Next, we compare other scores with the *MAA* metric with this combination of coefficients.

Table 2 shows the Precision and AUC values obtained for all proposed metrics, together with the theoretical bounds of the AUC. Interestingly, the AA_c score (classical *AA* metric for collaboration networks) has an AUC value quite close to the random one, but the second highest Precision after the *MAA* score. This reflects the fact that, even though the heuristic behind the AA_c metric seems to be a good proxy of the real dynamics, the limited amount of information hinders the prediction process. On the other hand, the keywords layer is the densest one and thus it carries much more information than the others, yielding a larger theoretical maximum for the AUC of the metrics based on this layer, such
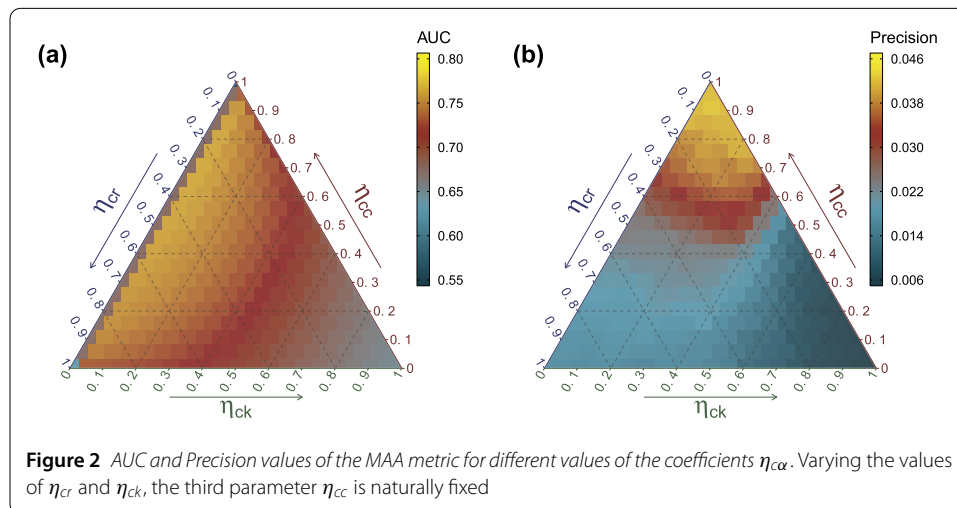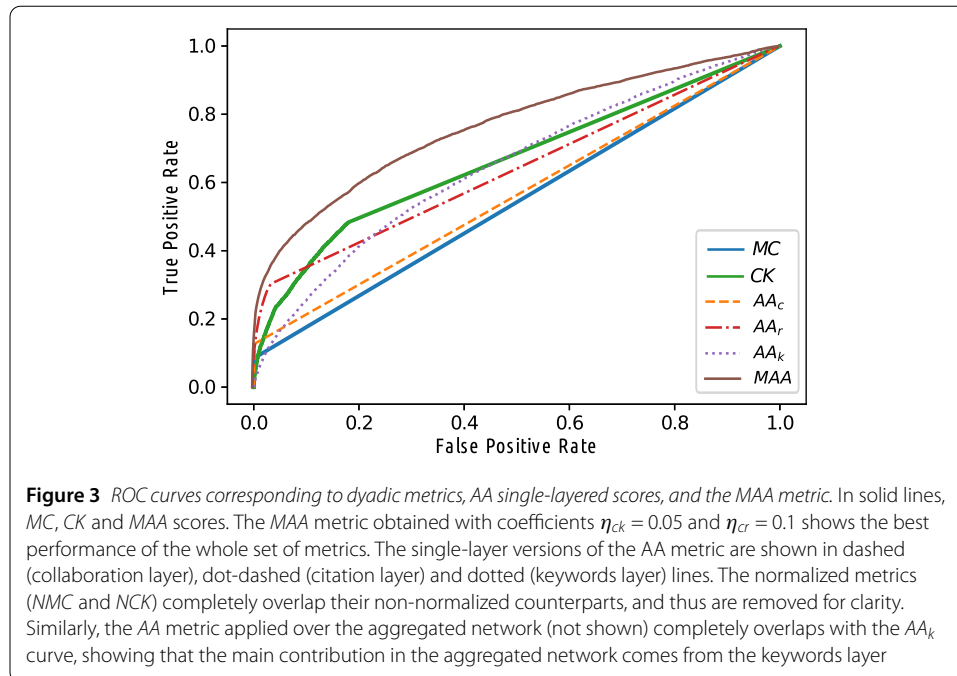

Figure 2 *AUC and Precision values of the MAA metric for different values of the coefficients $\eta_{c\alpha}$. Varying the values of η_{cr} and η_{ck}, the third parameter η_{cc} is naturally fixed*

Table 2 Precision and AUC values obtained for different metrics proposed, with the theoretical bounds of the AUC. We consider dyadic metrics given by Eqs (1)–(4), triadic closure given by the *AA* metric, Eq. (5), applied to each layer (AA_c, AA_r, and AA_k), to the aggregated network (AA_a), and the *MAA* score given by Eq. (6), with coefficients $\eta_{ck} = 0.05$ and $\eta_{cr} = 0.1$ which maximize both AUC and Precision (see Fig. 2). Note that dyadic (*MC*, *NMC*, *CK*, and *NCK*) and triadic (based on *AA*) methods use different amount of information, so the theoretical bounds for the AUC are different

Method	Precision	AUC	AUC [worst–best]
Random	$4.3 \cdot 10^{-4}$	0.5	[0.50–0.50]
MC	0.025	0.5421	[0.5420–0.5422]
NMC	0.023	0.5422	[0.5420–0.5422]
CK	0.012	0.6648	[0.6082–0.6952]
NCK	0.005	0.6618	[0.6082–0.6952]
AA_c	0.041	0.5635	[0.5633–0.5636]
AA_r	0.017	0.6361	[0.6282–0.6393]
AA_k	0.006	0.6481	[0.0171–0.9951]
AA_a	0.006	0.6495	[0.0147–0.9971]
MAA (all triads)	0.042	0.7620	[0.0147–0.9971]

as the *CK*, *NCK*, and AA_k scores. However, the Precision of these scores is not as good as other metrics, indicating that sharing keywords is not such a good descriptor of the dynamics behind establishing new collaborations. Note that the *AA* score applied to the aggregated, single-layered network given by the projection of all layers onto a single layer (AA_a) is indistinguishable from the AA_k score, given that the projected network is dominated by the keywords layer. Metrics based on the citation layer show a behavior between the other two: the citation layer carries less information than the keyword layer but more than the collaboration one. Therefore, the AUC value of the AA_r method is larger than the AUC of the AA_c. Note, however, that dyadic metrics such as the *MC* and *NMC* scores have a much lower AUC (also lower than the AA_c score), even if they show a slightly larger Precision. This indicates that simple dyadic metrics cannot outperform scores based on triadic closures with respect to citations. Finally, the *MAA* metric given by Equation (6) with coefficients $\eta_{ck} = 0.05$ and $\eta_{cr} = 0.1$, which maximize both AUC and Precision, has a much larger AUC and Precision than all other single layered metrics.

The detailed ROC of different dyadic and triadic scores are showed in Fig. 3. Note that the curves obtained by normalized scores (NMC and NCK) are not shown, since they exactly overlap the corresponding ROC for non-normalized scores (MC and CK, respectively), indicating that the normalization factor has no effect. Also, the ROC obtained by applying the *AA* score to the aggregated network, AA_a, is equivalent to the ROC of the AA_k score, and thus it is not shown, for clarity. This behavior is also confirmed by the AUC values reported in Table 2. Figure 3 unveils that the ROC curve of the *MAA* metric given by Eq. (6), with coefficients $\eta_{c\alpha}$ that maximize the AUC, clearly outperforms all other metrics. Finally, Fig. 3 clearly shows the point of the ranking beyond which only scoreless links remain, and thus the curves start to follow a linear trend. This point is different for different metrics and it is responsible for the theoretical bounds of the AUC indicated in Table 2. Consistently with this behavior, the only curves that do not show such inflection points are the ones corresponding to the metrics with no theoretical bounds on the AUC, namely the AA_k, AA_a, and *MAA* scores.

Figure 3 *ROC curves corresponding to dyadic metrics, AA single-layered scores, and the MAA metric.* In solid lines, *MC*, *CK* and *MAA* scores. The *MAA* metric obtained with coefficients $\eta_{ck} = 0.05$ and $\eta_{cr} = 0.1$ shows the best performance of the whole set of metrics. The single-layer versions of the AA metric are shown in dashed (collaboration layer), dot-dashed (citation layer) and dotted (keywords layer) lines. The normalized metrics (*NMC* and *NCK*) completely overlap their non-normalized counterparts, and thus are removed for clarity. Similarly, the *AA* metric applied over the aggregated network (not shown) completely overlaps with the AA_k curve, showing that the main contribution in the aggregated network comes from the keywords layer

5 Conclusions

To sum up, we have shown that scientific credit and common scientific interests can be predictive of new collaborations between scientists. For this purpose, we reconstructed a dataset of publication records by merging different bibliographic sources, including keywords that indicate the topics of papers. We represent this dataset as a multiplex network, in which each layer encodes a different kind of interaction, directed or undirected. Next, we compare several link prediction algorithms, based on different dyadic and triadic interactions between scientists. Our findings show that metrics based on triadic closure generally outperform simpler dyadic scores, and that the contributions of different layers are bounded by the amount of information available in each layer. For this reason, the best results, both in terms of Precision and AUC, are obtained by combining the information present in different layers by means of the Multiplex Adamic-Adar score [31], that fully exploits the multiplex nature of the scientific networks reconstructed here.

The coefficients that maximize the Multiplex Adamic-Adar metric indicate how the information structured in the multiplex network can be optimized for the prediction of new scientific collaborations. In this regard, one can notice that the major contribution is given, as expected, by the collaboration layer, while contributions from citation and keywords layers are smaller. For the keyword layer, this is due to its large density, that improves AUC but may reduce Precision. A possible improvement to the prediction of new collaborations could thus be given by a smaller, more precise set of keywords able to better map the different fields of Physics [8]. In future works, it would be interesting to incorporate additional information from publication records into the scientific multiplex network, such as institutional affiliations and geographical locations, to see if these features are predictive of new collaborations. While in this work we predict new collaborations only within the field of Physics due to the computational complexity of link prediction algorithms, the dataset presented and the prediction metrics proposed here can be applied beyond the

Physics field, which will enable getting further insights and a better understanding of the differences across scientific fields.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-021-00282-x.

> **Additional file 1.** Supplementary information (PDF 272 kB)

Availability of data and materials
The data sets used and/or analysed during the current study are available from the corresponding author upon request.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MT and AA analyzed data and performed simulations. MS and DP designed research with contribution from AA and YM. All authors discussed the results and wrote the manuscript. All authors read and approved the final version of the manuscript.

Author details
[1]ISI Foundation, via Chisola 5, 10126 Torino, Italy. [2]Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, 50018 Zaragoza, Spain. [3]Department of Theoretical Physics, University of Zaragoza, 50018 Zaragoza, Spain.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
1. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, Petersen AM, Radicchi F, Sinatra R, Uzzi B et al (2018) Science of science. Science 359(6379):0185
2. Clauset A, Larremore DB, Sinatra R (2017) Data-driven predictions in the science of science. Science 355(6324):477–480
3. Sinatra R, Deville P, Szell M, Wang D, Barabási A-L (2015) A century of physics. Nat Phys 11(10):791
4. Battiston F, Musciotto F, Wang D, Barabási A-L, Szell M, Sinatra R (2019) Taking census of physics. Nat Rev Phys 1(1):89
5. Wagner CS, Roessner JD, Bobb K, Klein JT, Boyack KW, Keyton J, Rafols I, Börner K (2011) Approaches to understanding and measuring interdisciplinary scientific research (IDR): a review of the literature. J Informetr 5:14–26
6. Leydesdorff L, Rafols I (2011) Indicators of the interdisciplinarity of journals: diversity, centrality, and citations. J Informetr 5(1):87–100. https://doi.org/10.1016/j.joi.2010.09.002
7. Foster JG, Rzhetsky A, Evans JA (2015) Tradition and innovation in scientists' research strategies. Am Sociol Rev 80(5):875–908. https://doi.org/10.1177/0003122415601618
8. Aleta A, Meloni S, Perra N, Moreno Y (2019) Explore with caution: mapping the evolution of scientific interest in physics. EPJ Data Sci 8(1):27. https://doi.org/10.1140/epjds/s13688-019-0205-9
9. Bornmann L, Leydesdorff L, Walch-Solimena C, Ettl C (2011) Mapping excellence in the geography of science: an approach based on scopus data. J Informetr 5(4):537–546
10. Zhang Q, Perra N, Gonçalves B, Ciulla F, Vespignani A (2013) Characterizing scientific production and consumption in physics. Sci Rep 3:1640
11. Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabási A-L (2014) Career on the move: geography, stratification, and scientific impact. Sci Rep 4:4770
12. Sinatra R, Wang D, Deville P, Song C, Barabási A-L (2016) Quantifying the evolution of individual scientific impact. Science 354(6312):5239
13. Newman MEJ (2010) Networks. An introduction. Oxford University Press, London
14. Newman MEJ (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci USA 98(2):404–409. https://doi.org/10.1073/pnas.98.2.404
15. Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings of the twelfth international conference on information and knowledge management. CIKM '03. Association for Computing Machinery, New York, pp 556–559. https://doi.org/10.1145/956863.956972

16. Anand K, van Lelyveld I, Banai Á, Friedrich S, Garratt R, Hałaj G, Fique J, Hansen I, Jaramillo SM, Lee H, Molina-Borboa JL, Nobili S, Rajan S, Salakhova D, Silva TC, Silvestri L, de Souza SRS (2018) The missing links: a global study on uncovering financial network structures from partial data. J Financ Stab 35:107–119. https://doi.org/10.1016/j.jfs.2017.05.012

17. Wang P, Xu B, Wu Y, Zhou X (2015) Link prediction in social networks: the state-of-the-art. Sci China Inf Sci 58(1):1–38. https://doi.org/10.1007/s11432-014-5237-y

18. Lü L, Jin C-H, Zhou T (2009) Similarity index based on local paths for link prediction of complex networks. Phys Rev E 80:046122. https://doi.org/10.1103/PhysRevE.80.046122

19. Adamic LA, Adar E (2003) Friends and neighbors on the web. Soc Netw 25(3):211–230. https://doi.org/10.1016/S0378-8733(03)00009-1

20. Yao L, Wang L, Pan L, Yao K (2016) Link prediction based on common-neighbors for dynamic social network. Proc Comput Sci 83:82–89. https://doi.org/10.1016/j.procs.2016.04.102

21. Newman MEJ (2001) Clustering and preferential attachment in growing networks. Phys Rev E 64:025102. https://doi.org/10.1103/PhysRevE.64.025102

22. Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453(7191):98–101. https://doi.org/10.1038/nature06830

23. Cho H, Yu Y (2018) Link prediction for interdisciplinary collaboration via co-authorship network. Soc Netw Anal Min 8(1):1–12. https://doi.org/10.1007/s13278-018-0501-6

24. Al Hasan M, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In: SDM06: workshop on link analysis, counter-terrorism and security, vol 30, pp 798–805

25. Moradabadi B, Meybodi MR (2017) A novel time series link prediction method: learning automata approach. Physica A 482:422–432. https://doi.org/10.1016/j.physa.2017.04.019

26. Najari S, Salehi M, Ranjbar V, Jalili M (2019) Link prediction in multiplex networks based on interlayer similarity. Physica A 536:120978. https://doi.org/10.1016/j.physa.2019.04.214

27. Sett N, Ranbir Singh S, Nandi S (2016) Influence of edge weight on node proximity based link prediction methods: an empirical analysis. Neurocomputing 172:71–83. https://doi.org/10.1016/j.neucom.2014.11.089

28. Zhang J (2017) Uncovering mechanisms of co-authorship evolution by multirelations-based link prediction. Inf Process Manag Int J 53:42–51

29. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Netw 2(3):203–271. https://doi.org/10.1093/comnet/cnu016

30. Aleta A, Moreno Y (2019) Multilayer networks in a nutshell. Annu Rev Condens Matter Phys 10(1):45–62. https://doi.org/10.1146/annurev-conmatphys-031218-013259

31. Aleta A, Tuninetti M, Paolotti D, Moreno Y, Starnini M (2020) Link prediction in multiplex networks via triadic closure. Phys Rev Res 2(4):042029. https://doi.org/10.1103/PhysRevResearch.2.042029

32. Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80(5):056103. https://doi.org/10.1103/PhysRevE.80.056103

33. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: Proceedings of the fourteenth ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD '2008), pp 990–998

34. Menichetti G, Remondini D, Panzarasa P, Mondragón RJ, Bianconi G (2014) Weighted multiplex networks. PLoS ONE 9(6):e97857. https://doi.org/10.1371/journal.pone.0097857

35. Jia T, Ran Y, Xu X (2020) The bounds of similarity-based link prediction by the AUC measure. In: NetSci-X 2020