Supplementary material for "Evaluating the impact of interdisciplinary research: a multilayer network approach"

Elisa Omodei, Manlio De Domenico and Alex Arenas∗
Department of Mathematics and Computer Science, Rovira i Virgili University
(*e-mail:* elisa.omodei@urv.cat)

1 Comparison with other approaches

The idea of ranking scholars simulating a diffusion process was already introduced in (Radicchi *et al.*, 2009) – and previously in (Walker *et al.*, 2007) to rank scientific publications – but the approach proposed in this work considers a different kind of network – a bipartite interconnected multilayer network. In this section we motivate the choice of taking into account the complete bipartite structure instead of its one-mode projection.

For the sake of simplicity, we will consider a bilayer version of the network. Our focus here is not in fact the multilayer aspect of the network capturing interdisciplinarity, but rather its bipartition. Therefore, let us consider $N = N_P + N_A$ nodes and 2 layers $\{l_1, l_2\}$. The 4 components of the rank-2 adjacency tensor $C^{\alpha}_{\beta}(\tilde{h}, \tilde{k})$ are now defined as follows. $C^{\alpha}_{\beta}(\tilde{l}_1, \tilde{l}_1)$ encodes information about citing relations between papers, *i.e.* $w_{ij}(\tilde{l}_1, \tilde{l}_1) = 1$ if paper i cites paper j. $C^{\alpha}_{\beta}(\tilde{l}_1, \tilde{l}_2)$ encodes information about paper authorship, *i.e.* $w_{ij}(\tilde{l}_1, \tilde{l}_2) = 1$ if author j is one of the authors of paper i. Finally, we define $C^{\alpha}_{\beta}(\tilde{l}_2, \tilde{l}_1)$ and $C^{\alpha}_{\beta}(\tilde{l}_2, \tilde{l}_2)$ to be zero tensors, consistently with the representation introduced in S.1. For the sake of simplicity, since in the rest of the section we will be dealing only with rank-2 tensors, we will make use of the simpler classical matrix notation instead of the tensorial one. Therefore we will denote $C^{\alpha}_{\beta}(\tilde{l}_1, \tilde{l}_1)$ as C and $C^{\alpha}_{\beta}(\tilde{l}_1, \tilde{l}_2)$ as A.

In the author citation network proposed in (Radicchi *et al.*, 2009), each node represents an author, and $w_{ij} \neq 0$ if there exist at least one publication α, of which i is an author, that cites a publication β of which j is an author. Each such publication gives a contribution $\frac{1}{nm}$ to w_{ij} (where n is the number of authors of publication α, and m is the number of authors of β) so that the total contribution of each citation is equal to 1.

Let us consider an adjacency matrix $\mathscr{C}$ of size $N_P \times N_P$, encoding the citation links between papers. C can be built from $\mathscr{C}$ by means of multiplication with a rectangular matrix $\mathscr{I}$ of size $(N_P + N_A) \times N_P$ such that $(\mathscr{I})_{ii} = 1$ for $i = 1, \ldots, N_P$, and all the other

∗ **Acknowledgements:** A.A., E.O. and M.D.D. acknowledge financial support by the Spanish government through grant FIS2015-38266. M.D.D. also acknowledges financial support from the Spanish program Juan de la Cierva (IJCI-2014-20225). A.A. also acknowledges partial financial support from the European Commission FET-Proactive project MULTIPLEX (Grant No. 317532), ICREA Academia and James S. McDonnell Foundation.

elements are equal to 0. Then

$$C = \mathscr{I}\mathscr{C}\mathscr{I}^T. \tag{1}$$

Using $\mathscr{I}$ we can also build A from the projection matrix $\mathscr{P}$, of size $N_P \times N_A$, where $w_{ij} = 1$ if j is one of the authors of paper i:

$$A = \mathscr{I}\mathscr{P}\mathscr{I}^T. \tag{2}$$

Let us define $\tilde{\mathscr{P}}$ as the normalised version of $\mathscr{P}$, *i.e.* $(\tilde{\mathscr{P}})_{ij} = \frac{(\mathscr{P})_{ij}}{\sum_{k=1}^{N_A}(\mathscr{P})_{ik}} = \frac{1}{m_i}$ (where m_i is the number of authors of paper i), then the $N_A \times N_A$ adjacency matrix representing the network of citations between authors can be obtained performing two successive matrix multiplications:

$$\mathscr{A} = \tilde{\mathscr{P}}^T\mathscr{C}\tilde{\mathscr{P}}. \tag{3}$$

Proof:

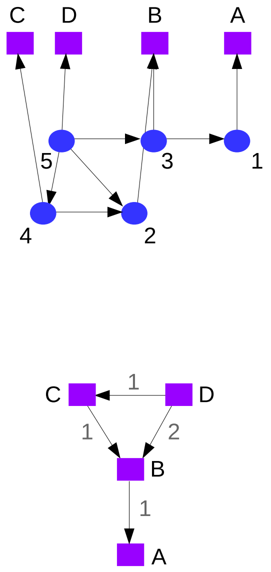
$$(\tilde{\mathscr{P}}^T\mathscr{C})_{ik} = \sum_{h=1}^{N_P}(\mathscr{P})_{hi}(\mathscr{C})_{hk} = \sum_{\substack{h=1\\(\mathscr{P})_{hi}\neq 0,(\mathscr{C})_{hk}=1}}^{N_P} \frac{1}{m_h} \tag{4}$$

This means that $(\tilde{\mathscr{P}}^T\mathscr{C})_{ik}$ is a sum over the papers authored by i that cite paper k, where each paper h gives a contribution of 1 over the number of authors. Then:

$$(\mathscr{A})_{ij} = \sum_{k=1}^{N_P}(\tilde{\mathscr{P}}^T\mathscr{C})_{ik}(\mathscr{P})_{kj} = \sum_{\substack{k=1\\(\mathscr{P})_{kj}\neq 0}}^{N_P} \Bigg(\sum_{\substack{h=1\\(\mathscr{P})_{hi}\neq 0,(\mathscr{C})_{hk}=1}}^{N_P} \frac{1}{m_h}\Bigg)\frac{1}{m_k} \tag{5}$$

Each element $(\mathscr{A})_{ij}$ is therefore a sum over all the pairs of papers (h,k) such that i is an author of h and j of k, and each element of the sum gives a contribution equal to $\frac{1}{m_h m_k}$, as indeed defined in (Radicchi *et al.*, 2009). □

We have demonstrated that the adjacency matrix of the author citation network is obtained performing two operations of matrix multiplication involving $\mathscr{C}$ and $\mathscr{P}$. Matrix multiplications consists in multiplications and summations of the matrix elements, which inevitably lead to information loss. The supra-adjacency matrix of the network proposed in this paper is instead the sum of the two expansions of $\mathscr{C}$ and $\mathscr{P}$, *i.e.* C and A, respectively. This guarantees that no information is loss, and this why we chose to consider the whole bipartite structure. Figure 1 shows an example in which the information loss characteristic of the author citation network leads to a less fair ranking of authors compared to the ranking based on the network introduced in this paper. Using our approach (top figure), the most central author is B, who is the author of both the most central paper and of one of the second most central papers. However, in the author citation network framework, the most central author is A (to understand why, we recall that the PageRank centrality is based not only on the number of incoming edges, but on the importance of the nodes from which these edges originate). This is due to the fact that in the author citation network all the information about an author's incoming citations from different papers is aggregated, and therefore A benefits from the importance of B without any distinction between the importance coming from papers that actually cite A, and that coming from B's other papers. In this case B's most central (and cited) paper is not the one citing A's paper, but this information is lost



Ranking	
#1	B
#2	A
#3	C
#4	D

Ranking	
#1	A
#2	B
#3	C
#4	D

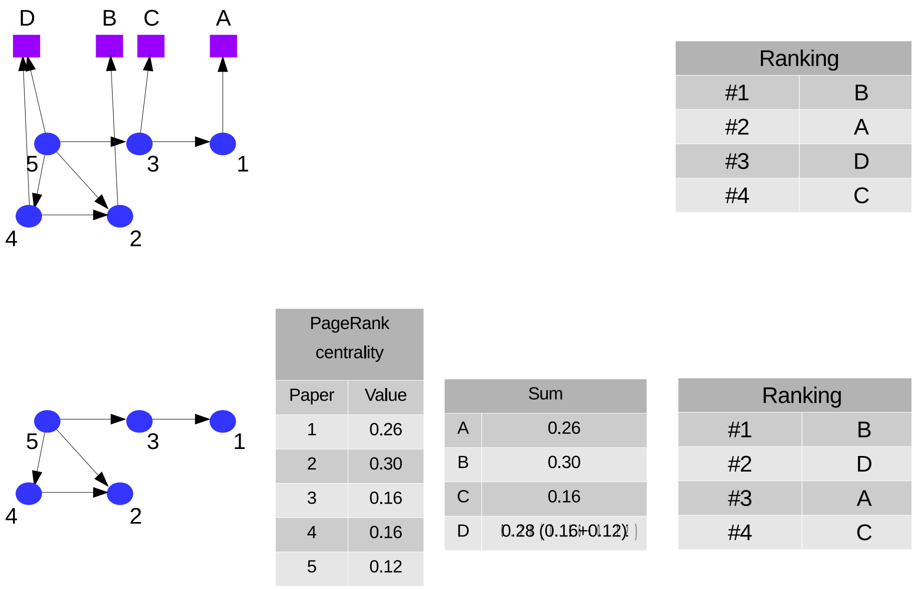
Fig. 1. An example in which the PageRank centralities computed on the author citation network and on the bipartite network lead to different rankings of the authors.

in the author citation network. As a consequence, the resulting ranking does not always reflect the real importance of the different authors.

An alternative approach is to use the PageRank method to get the centrality of papers in $\mathscr{C}$, and then compute the author centrality as a properly normalised sum of the centralities of the papers she/he has authored. In matrix terms, the author PageRank centrality vector ω_A can be obtained by simply applying a linear transformation to the paper PageRank centrality vector ω_P:

$$\omega_A = \tilde{\mathscr{P}}^T \omega_P. \tag{6}$$

However, this solution involves another kind of aggregation which can lead to misleading results too. An example is shown in Figure 2. Using the sum approach, author *D* becomes more central than author *A*, because she/he authored two papers, even though they are the two most marginal papers in the network (note that the only citation to paper 4 is a self-citation). On the contrary, *A* is the author of a very central paper, and in fact our approach correctly classifies her/him as more central than *D*. The issue with this alternative approach is that the PageRank is a diffusion process, which is not a linear dynamics. Therefore summing over the centralities of different nodes is also an aggregation process through which some information on the system is lost.



Ranking	
#1	B
#2	A
#3	D
#4	C

PageRank centrality	
Paper	Value
1	0.26
2	0.30
3	0.16
4	0.16
5	0.12

Sum	
A	0.26
B	0.30
C	0.16
D	0.28 (0.16+0.12)

Ranking	
#1	B
#2	D
#3	A
#4	C

Fig. 2. An example in which the PageRank centralities computed as the sum over the papers and on the bipartite network lead to different rankings of the authors.

2 Productivity control

In Figure 3 of the main text, we show that we find a strong positive correlation between the gain in rank that scholars and inventors obtain when evaluated using the proposed method – instead of a method based on a flat representation of the citation network –, and their topical interdisciplinarity (panels (a) and (b)). Moreover, we show that the rank gain is also positively correlated with the disciplinary diversity of scholars' and inventors' incoming citations (panels (c) and (d)). To control for the effects of productivity, i.e. the fact that a researcher that has produces more papers has more chances to publishes in more areas or to be cited by papers in many different areas, we perform the same analysis on two subsets of the data by considering in each case only authors with a fixed number of publications. Figure 3 shows the result for the subset of authors and inventors with 20 ± 2 publications, and Figure 4 for authors and inventors with 50 ± 2 publications. The results are consistent with those obtained using the whole dataset.

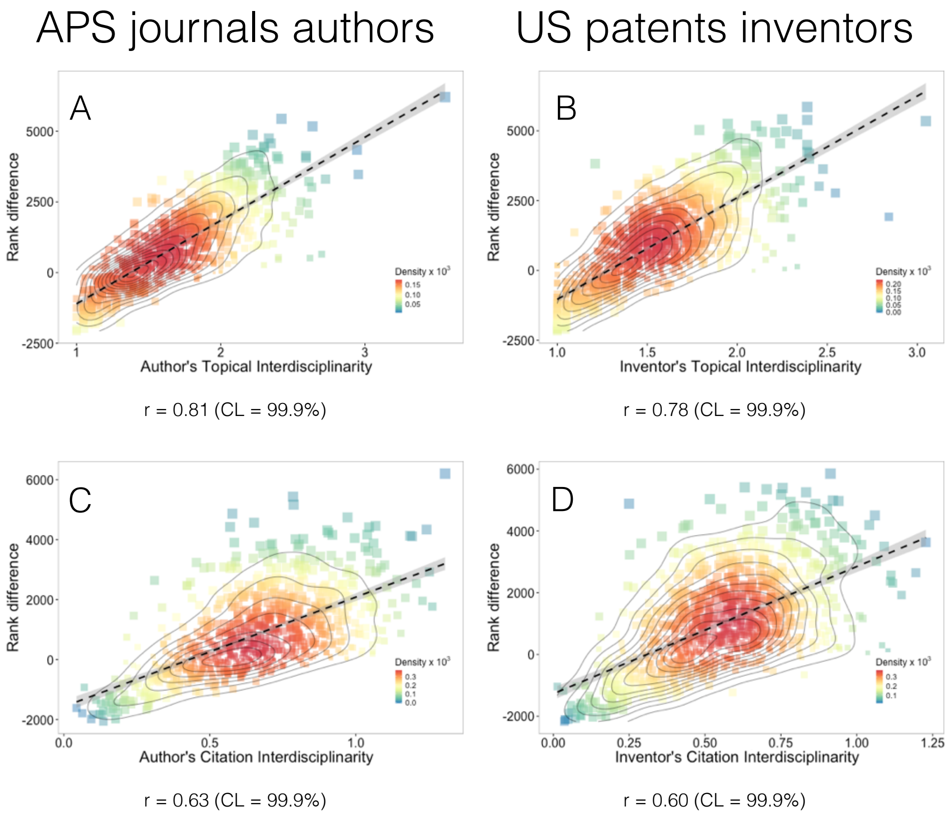


Fig. 3. **Correlations.** Heat-maps representing the correlation between the gain in rank that scholars and inventors with 20 ± 2 publications obtain when evaluated using the proposed method – instead of a method based on a flat representation of the citation network –, and two measures of their interdisciplinarity level. The x-axis represents, in panel (a) and (b), scholars' and inventors' topical interdisciplinarity, defined as the average number of different scientific areas their publications pertain to, and, in panel (c) and (d), their diversity in terms of disciplines of the scholars' and inventors' incoming citations (citation interdisciplinarity). Correlations are calculated using Pearson's r coefficient, and setting the statistical significance at 0.1%. Solid lines represent density gradient contours, and dashed lines represent linear regression models estimated via maximum-likelihood.

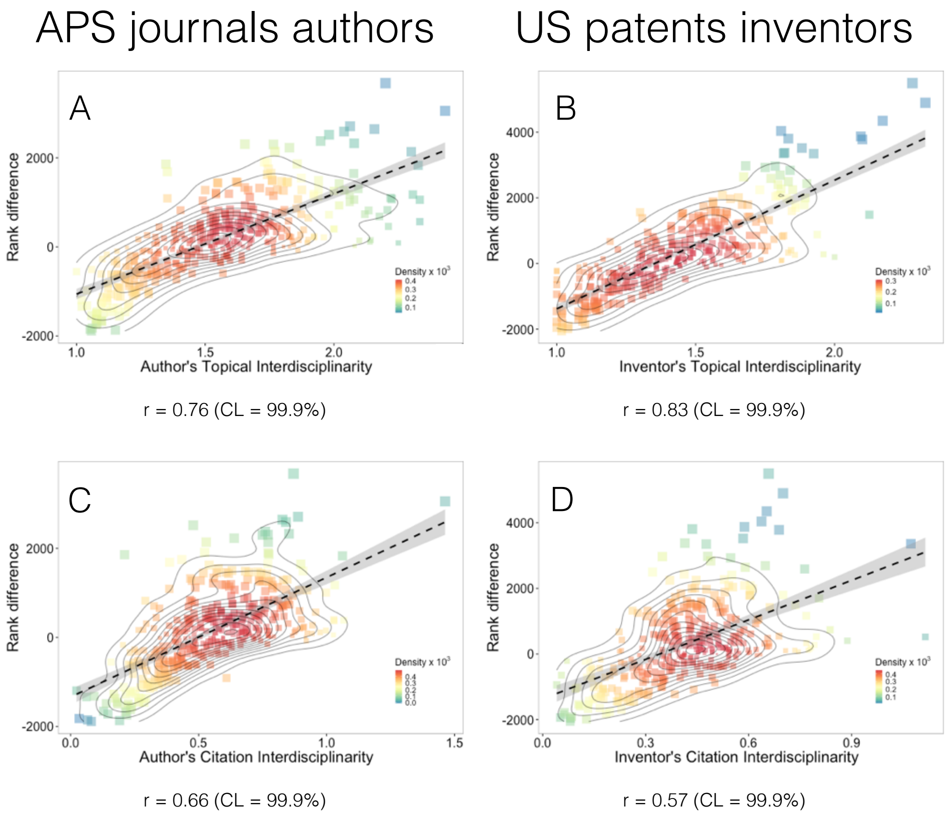


Fig. 4. **Correlations.** Heat-maps representing the correlation between the gain in rank that scholars and inventors with 50 ± 2 publications obtain when evaluated using the proposed method – instead of a method based on a flat representation of the citation network –, and two measures of their interdisciplinarity level. The x-axis represents, in panel (a) and (b), scholars' and inventors' topical interdisciplinarity, defined as the average number of different scientific areas their publications pertain to, and, in panel (c) and (d), their diversity in terms of disciplines of the scholars' and inventors' incoming citations (citation interdisciplinarity). Correlations are calculated using Pearson's r coefficient, and setting the statistical significance at 0.1%. Solid lines represent density gradient contours, and dashed lines represent linear regression models estimated via maximum-likelihood.

References

Radicchi, Filippo, Fortunato, Santo, Markines, Benjamin, & Vespignani, Alessandro. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical review e*, **80**(5), 056103.

Walker, Dylan, Xie, Huafeng, Yan, Koon-Kiu, & Maslov, Sergei. (2007). Ranking scientific publications using a model of network traffic. *Journal of statistical mechanics: Theory and experiment*, **2007**(06), P06010.