A Comparative Analysis of Interdisciplinary Co-Authorship Networks

Ortona Gaia, Digital Humanities and Digital Knowledge, 0001102809
Guenci Matteo, Digital Humanities and Digital Knowledge, 0001102177
Zilli Leonardo, Digital Humanities and Digital Knowledge, 0001102264

1 Introduction

The research described in this report focuses on the domain of scientometrics, an area of scholarly investigation concerned with the quantitative analysis of bibliographic information to assess diverse facets of academic publications. In particular, we set to study the co-authorship patterns among the academic research circles of three master's courses spanning diverse disciplines with the aim of analyzing the topological structure of the three courses through their respective teachers and their collaborative endeavors in the academic research.

2 Problem and Motivation

The idea for the research first emerged from the will to analyze the inner dynamics of our own master's course in Digital Humanities and Digital Knowledge (DHDK), whose program celebrates with pride its interdisciplinary nature that integrates insights from the humanities with the methodologies of computer science. Our aim was then to investigate its structure and its connection to the different disciplines that it teaches, and we did that by examining the academic output of its professors and leveraging the co-authorship network that emerged, which offered an ideal lens for this analysis.
To have a better grasp on the investigation we have chosen to extend our research with a comparative analysis between the co-authorship network of our course and those of the Computer Science and Italian Studies master's courses. This comparative study allowed us to discern whether the behavior of the academic collaborations within the DHDK network aligns more closely with those typical of "exact" sciences or those rooted in the "social" sciences.

3 Datasets

The dataset used for the analyses described in this report was collected by us through the extraction of the data coming from the IRIS, the online database of the CRIS[1] system, which collects and manages the outcome of University of Bologna's research activity. We have identified an initial pool of authors for each of the three master's programs by selecting the teacher of each

[1] https://cris.unibo.it/

one of the courses displayed on the "course structure diagram" webpage relative to the a.y. 22-23 of each master's website page[2][3][4] (for the Computer Science master's course the "Tecniche del software" curriculum was selected, and the "Italianistica" curriculum for the Italian Studies course).
We have then scraped the IRIS portal for the publications of each one of these teachers, retrieving for each publication the list of authors internal to the University of Bologna. Then, for each one of both the authors from the initial pool and their collaborators, we have retrieved, again through the IRIS website, the affiliation data of each.
It is important to specify that for the creation of the networks, only the first 19 professors with the most publications along with their co-authors have been inserted as nodes in an attempt to align the size of the network of the IT and CS courses (that had respectively 59 and 27 professors each) to the DHDK one.
This data has subsequently been inserted into a graph using the *NetworkX* python library to build three distinct undirected, weighted networks. More in detail, given two nodes u and v, the edge (u, v) represents the fact that researchers u and v have co-authored a publication. The nodes have been enriched with the affiliation data for each, describing the department under which the researcher works, and the edges have been given a weight denoting the number of co-authorship that have been found between any two connected nodes.

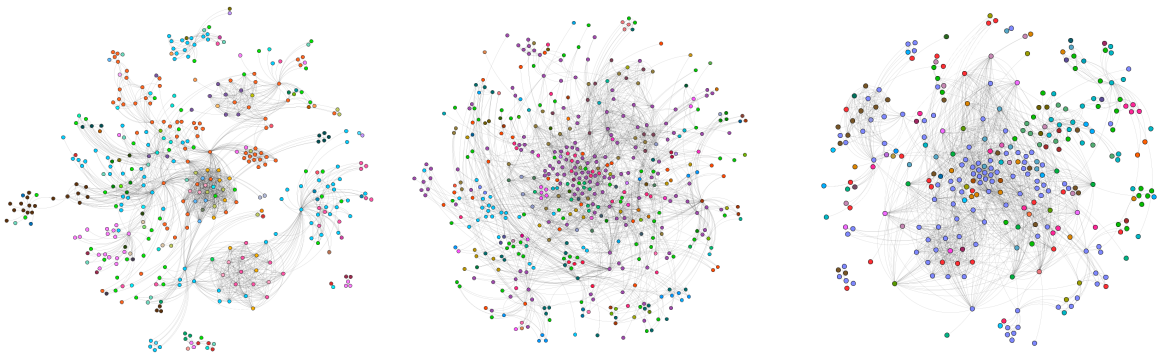

Figure 1: Visual representation of co-authorship networks for DHDK, CS, and IT. Color-coded on the affiliation attribute of each node and visualized through Gephi.

Table 1: Number of nodes and edges of each graph.

Network	***Number of nodes***	***Number of edges***
DHDK	398	1889
CS	493	1999
IT	304	1523

[2]https://corsi.unibo.it/2cycle/DigitalHumanitiesKnowledge/course-structure-diagram/piano/2023/9224/000/000/2022
[3]https://corsi.unibo.it/2cycle/ComputerScience/course-structure-diagram/piano/2022/5898/A58/000/2022
[4]https://corsi.unibo.it/magistrale/ItalianisticaCultureLetterarieLinguistica/insegnamenti/piano/2023/9220/A87/000/2022

4 Validity and Reliability

Being the dataset sourced directly from the records of the University of Bologna, we are confident of a certain degree of faithfulness. It has to be noted, however, that the records are autonomously updated by their authors, possibly introducing inconsistencies.
Regarding the collection of the data, the initial selection of teachers surely contributes in adding a "bias" to the networks, which are built from the point of view of a restricted number of nodes. The cutoff we set to the number of initial nodes also might have removed some precious data from the CS and IT datasets, but we deemed it necessary to standardize the size of the three networks.
Despite these problems, the same exact methodologies have been used in the collection of all three networks, which should give a solid ground to the comparative aspect of the research.

5 Measures and Results

5.1 Centrality measures

In order to evaluate the performance of each of the authors in the network, we evaluated micro-level indicators such as the centrality measures to investigate the most central and influential nodes within the networks, focusing on the contrast of the resulting values obtained from taking into consideration the unweighted edges first and then their weighted version.
The analysis revealed highly influential nodes critical to the network's robustness and resilience, significantly affecting its dispersion rate if removed. Specifically, Francesca Tomasi and Silvio Peroni emerged as influential nodes in DHDK based on the their degree centrality(though their centrality remains high in the IT master network), while Luciano Bononi and Tullio Cinotti Salmon represent influential nodes in CS. After establishing these nodes' high connectivity, we further examined the quality of their connections using eigenvector centrality, considering both the number and significance of a node's neighbors. Notably, in the DHDK network, Francesca Tomasi and Silvio Peroni retained high centrality values (0.198 and 0.197, respectively) in the unweighted network, indicating strong connections with well-connected nodes.
However, with weighted edges, the network's structure shifted, emphasizing that nodes with higher eigenvector centrality in the weighted network are influential not just due to their connectivity but also because of the strength of their connections (considering co-authorship frequency). Professors such as Monica Palmirani or Valentina Presutti, previously absent from the higher positions, now exhibit higher eigenvector values due to their connections with central nodes. The same goes for the other two networks.
Regarding betweenness and closeness centralities, we applied them to check the connectivity and compactness of the networks.
With closeness we focused on understanding how close a node is to all the other nodes, considering edges with and without weights. Higher values indicate nodes that are, on average, closer to the others.
The findings indicate that elevated edge weights in a network proportionally augment the inter-nodal distance. When traversing a path characterized by high edge weights, its contribution to the overall distance within the weighted network becomes more pronounced. The presence of high-weighted edges significantly influences the computation of distances between nodes, amplifying the perceived length or cost of paths in network analyses reliant on distance metrics. In the case of DHDK for example, Francesca Tomasi has the highest closeness centrality's value

without weights but with weighted edges her value decreases, proving what stated beforehand. Relating to betweeness centrality, nodes with higher values act as bridges, connecting different parts of the network. These individuals play a crucial role in maintaining efficient communication paths. We noticed that the shape of the networks changes with weighted edges, meaning that based on the co-authoring frequency, different nodes maintain the network compact.

Name	Degree Centrality	Betweenness Centrality		Closeness Centrality		Eigenvector Centrality	
		Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted
TOMASI, FRANCESCA	0.214	0.366	0.383	0.184	0.420	0.280	0.198
PERONI, SILVIO	0.189	0.098	0.185	0.150	0.395	0.539	0.197
MILANO, MICHELA	0.166	0.157	0.171	0.128	0.278	0.000018	0.00106
VITALI, FABIO	0.139	0.106	0.150	0.147	0.380	0.531	0.0395
BARTOLINI, ILARIA	0.123	0.310	0.310	0.160	0.355	0.00211	0.00872

Table 2: DHDK Centrality measures (weighted and unweighted). Table ordered by Degree centrality in descending order

Name	Degree Centrality	Betweenness Centrality		Closeness Centrality		Eigenvector Centrality	
		Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted
BONONI, LUCIANO	0.142276	0.082177	0.117018	0.210797	0.362564	0.5622057825	0.272
SALMON CINOTTI, TULLIO	0.119919	0.027431	0.041136	0.187571	0.334466	0.2193696974	0.264
MARFIA, GUSTAVO	0.119919	0.154433	0.172569	0.214379	0.360968	0.0104894254	0.0355
VITALI, FABIO	0.111789	0.133884	0.203039	0.212987	0.368263	0.0008061745	0.006
DI FELICE, MARCO	0.109756	0.039118	0.073842	0.164000	0.327563	0.5688782699	0.170

Table 3: CS Centrality measures (weighted and unweighted). Table ordered by Degree centrality in descending order

Name	Degree Centrality	Betweenness Centrality		Closeness Centrality		Eigenvector Centrality	
		Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted
TOMASI, FRANCESCA	0.280528	0.370793	0.490807	0.180190	0.329182	0.3389487662	0.198
CHINES, LOREDANA	0.148515	0.042855	0.069310	0.176548	0.305437	0.0083982820	0.0124
DAQUINO, MARILENA	0.135314	0.009556	0.003716	0.146399	0.259735	0.0376362767	0.184
TINTI, PAOLO	0.128713	0.077326	0.113041	0.151641	0.302135	0.0055537653	0.0121
PERONI, SILVIO	0.128713	0.019465	0.049502	0.150094	0.269244	0.0277595983	0.183

Table 4: IT Centrality measures (weighted and unweighted). Table ordered by Degree centrality in descending order

5.2 Clustering, K-core and Community detection

The **Clustering Coefficient** measures the extent to which nodes in a network tend to cluster together. A higher clustering coefficient suggests a network with dense local connections, highlighting cohesive subgroups. It provides insights into the network's resilience and the likelihood of interconnected nodes forming clusters. Nodes that have non-zero clustering coefficients with edge weights indicate that their neighbors are more likely to be connected to each other.
Nodes that have zero clustering coefficients with and without edge weights suggest that their neighbors are less likely to form connections among themselves. Local Clustering Coefficient (C_u) for node u is:

$$C_u = \frac{2 \cdot T(u)}{\deg(u) \cdot (\deg(u) - 1)}$$

Here, $T(u)$ represents the number of triangles that node u is part of, and $\deg(u)$ is the degree of node u. The clustering coefficient C_u provides a measure of how interconnected the neighbors of a node are. The change in clustering coefficients when considering edge weights is a result of

the weights influencing the likelihood and strength of connections between neighboring nodes.

The **Greedy Modularity Algorithm** is a community detection method that iteratively optimizes the modularity of a network to reveal meaningful node groupings. It measures the quality of network division, emphasizing dense connections within groups and sparse connections between them. The algorithm starts by assigning each node to its own community and iteratively merges communities to enhance overall modularity. The algorithm evaluates potential modularity gain from merging communities, selecting pairs that maximize it. Then actual vs. expected edges within communities are computed with the aim to maximize the sum of contributions across all communities.
The "greedy" nature lies in locally optimal choices of merging communities based on immediate modularity gains. The result is a partition with nodes having stronger connections within their communities. This process continues until further mergers no longer improve modularity.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

The results provided in the different tables are a mapping of nodes to their assigned communities, represented by numeric labels. Each node is associated with the community to which it belongs. Rranging from -1 to 1, values closer to 1 reflect a more cohesive community structure, while those around 0 suggest a network division similar to randomness. Negative values imply a less effective community structure. Weighted modularity considers both edge presence and strength, significantly impacting community detection, particularly when nodes share strong weighted connections.

Network	***Modularity***
DHDK	0.668
CS	0.769
IT	0.698

Table 5: Modularity values for DHDK, CS and IT networks

K-Core analysis reveals a network's core structure by identifying maximal sub-graphs (k-cores) where each node has a degree of at least k. Nodes with degrees less than k are iteratively removed, exposing increasingly connected structures. Nodes are assigned K-Core values, with higher values indicating stronger connectivity within the core. This method unveils central, well-connected nodes, providing insights into the network's core-periphery organization. The formula for obtaining the k-core sub-graph, denoted as $G(k)$, is as follows:

$$G(k) = G \setminus \{v \in G : \text{degree}(v) < k\}$$

Here, G represents the original graph, and the operation $\setminus$ denotes the removal of nodes with a degree that is less than k. This process is repeated at each iteration for increasing values of k until convergence, resulting in a series of k-core sub-graphs. This involves the `k_core` function in NetworkX, which directly computes the k-core sub-graph for a given value of k.
The **K-Core Value** (k_v) for node v:

$$k_v = \max_k \{k : v \in G^{(k)}\}$$

Here, $G^{(k)}$ represents the k-core sub-graph, and k_v is the K-Core value assigned to node v. This value represents the highest k for which node v remains part of the k-core sub-graph. This formula is related to the `core_number` function in NetworkX that calculates the core number for each node, representing the largest k-core to which the node belongs.
This provides a numerical representation of a node's k-core centrality. The choice to employ the `core_number` function in this analysis was taken because it provides a practical means to obtain numerical data, representing the k-core structure of each node in the network, to enhance the interpretability of the K-Core analysis. The presence of high and low core numbers suggests the existence of hierarchical or modular structures in the network:

- The hierarchical structure is evident in the tiered distribution of K-Core values. Nodes with high core numbers form a cohesive core, connecting different parts of the network, while nodes with low core numbers are often on the outskirts, serving more specialized roles.
- Simultaneously, the repeated occurrence of the same K-Core value for multiple nodes suggests the existence of distinct modules, contributing to a modular network structure.

Name	Clustering		Kcore	Community	
	Weighted	Unweighted		Weighted	Unweighted
MAMBELLI, FRANCESCA	0.075	0.800	4	2	0
SCOPECE, FIORA	0.074	1	9	6	8
DIONIGI, IVANO	0.061	0.866	12	6	8
CITTI, FRANCESCO	0.060	0.850	12	6	8
NERI, CAMILLO	0.059	0.850	12	6	8

Table 6: DHDK Clustering measure, KCore, and Community detection. Table ordered by Weighted Clustering in descending order

Name	Clustering		Kcore	Community	
	Weighted	Unweighted		Weighted	Unweighted
MONTALI, MARCO	0.108	1	4	7	6
LUCCHI, ROBERTO	0.091	1	5	0	1
MONTECCHIARI, LEONARDO	0.071	1	3	2	2
BRACUTO, MICHELE	0.064	1	3	2	3
ROUHI, RAHIMEH	0.064	1	2	5	8

Table 7: CS Clustering measure, KCore, and Community detection.
Table ordered by Weighted Clustering in descending order

Name	Clustering		Kcore	Community	
	Weighted	Unweighted		Weighted	Unweighted
CITTI, FRANCESCO	0.261	0.85	12	6	3
PASETTI, LUCIA	0.260	0.85	12	6	3
ZIOSI, ANTONIO	0.260	0.85	12	6	3
NERI, CAMILLO	0.260	0.85	12	6	3
PIERI, BRUNA	0.260	0.85	12	6	3

Table 8: IT Clustering measure, KCore, and Community detection. Table ordered by Weighted Clustering in descending order

5.3 Homophily measures

We have used the *affiliation* attribute of each node to compute the assortativity coefficient r, representing the tendency of nodes in the network to connect with other nodes that have similar attributes to them.

Network	**r**
DHDK	0.2374
CS	0.1840
IT	0.1584

Table 9: Assortativity coefficient of each networks

The obtained results tell us that in all three networks the propensity of the authors to strictly collaborate with authors with similar attributes is quite low. However, the assortativity of the DHDK network is slightly larger than the one of the two other networks, implying a stronger inclination of the teachers of this course to write a publication with colleagues coming from their very same department.

5.4 Structural analyses

5.4.1 Scale freedom

Inspecting a plot of the degree distribution of each network reveals that a large percentage of nodes have a low degree compared to a smaller pool of nodes that have a rather high degree, outlined in the histograms by the taller characteristic "tail" of the distribution. This suggests that the networks we are dealing with might fall into the category of *scale-free* networks, that is networks whose degree distribution follows a power-law behaviour. These types of networks are found to be highly robust, which would mean that they would survive the removal of even a sensible number of their nodes. To further investigate this possibility, a log-log plot of the degree distributions of the networks is produced and observed. This new scale should reduce the noise in our plot and help us identify a power-law by depicting the distribution following a straight-line behaviour. Clearly that is not what can be seen in our plots, however a visual analysis of the distribution can only take us so far in detecting scale freedom and more statistically accurate measurements need to be performed.

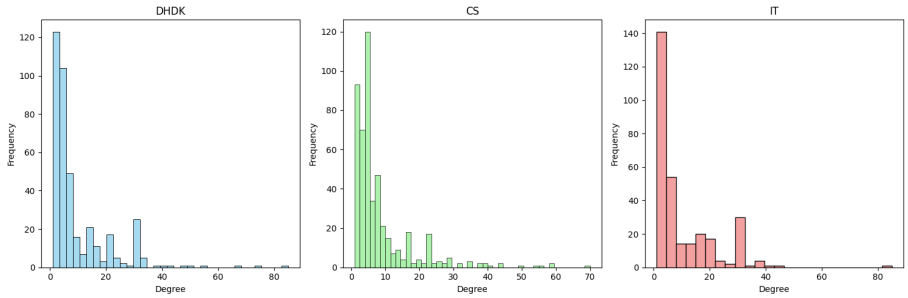

Figure 2: Degree distribution plot of the networks. From left to right: DHDK, CS, IT

To be absolutely certain that we are not dealing with scale-free networks, we can compute the scaling coefficient α as follows:

$$\alpha = 1 + n(\sum_i ln\frac{d_i}{d_{min} - 1/2})^{-1}$$

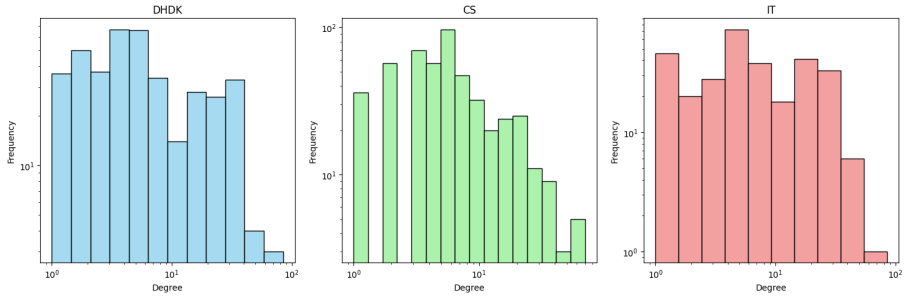

Figure 3: Degree distribution plot of the networks on a log-log scale. From left to right: DHDK, CS, IT

Where n is the number of nodes in our network, d_i is the degree of the node i and d_{min} the minimum degree found in the graph.
Finding an α value such that $2 \leq \alpha \leq 3$ would mean that the network is a scale-free network.

Table 10: α coefficient of each network.

Network	α
DHDK	1.414
CS	1.425
IT	1.411

As observed by the calculation of the α coefficient of our networks, we can then conclude with more certainty that neither of the three networks' degree distributions follow a power-law distribution, denying the hypothesis that they might be scale-free networks and thus suggesting that they are not as robust as a scale-free network would be.

5.4.2 Small-worldness

To further investigate the structural robustness of the networks we set to assess whether any of the networks presented a *small-world effect*. A small-world network is, simply put, a network that has shorter-than-expected distances between pairs of nodes. This phenomenon can be descriptive of different properties of networks, such as the presence of highly-clustered groups, "mediator" hubs and a particular robustness to random perturbation.
In order to compute the probabilities of our networks to display small-world features two calculations have been evaluated:

- the measure σ, obtained by comparing the clustering coefficient and average shortest path length of a network to the same measures obtained from an equivalent random network. The likelihood that a network presents a small-world configuration is found to be positive if $\sigma > 1$. It is worth noting that this measure is found to be particularly sensible to the size and density of the network, thus prompting us to compute a second measure.
  It is defined as:
$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$
- the measure ω, obtained by comparing the clustering coefficient C and average shortest path length L of a network to the clustering coefficient of an equivalent regular lattice

network C_l and to the average shortest path length of an equivalent random network L_r, respectively. Values such as $\omega < 0$ suggest that the network resembles a regular one, whereas for $\omega > 0$ the network is more akin to a random graph. $\omega \approx 0$ suggest a strong likelihood that the network presents small-world characteristics.
It is defined as:

$$\omega = \frac{L}{L_r} - \frac{C}{C_l}$$

Table 11: Average clustering coefficient, Average shortest path length, σ and ω of the three networks.

Network	***C***	***L***	σ	ω
DHDK	0.783	2.056	45.723	−0.158
CS	0.761	3.979	33.007	−0.721
IT	0.726	3.020	20.282	−0.550

The results obtained from the computations of the coefficients show that all three networks have a similarly high average clustering coefficient C, with DHDK being the most clustered network, suggesting that in all three courses the authors are more likely to form clusters of collaborations following, unsurprisingly, non-random patterns.
On the other hand, the three network present different values for what concerns the average shortest path length L, indicating that the nodes in the DHDK network are, on average, quite more close among each other compared to the nodes in the IT network and even more compared to the CS network.
C and L have then been used to compute the σ and ω measures, which yielded interesting results. All three networks appear to carry a very strong likelyhood to be small-world networks. The value of σ is extremely high in all three cases, but in particular the DHDK and CS networks both show exceptionally large coefficients.
The value of ω seems to confirm the estimations of σ, with the DHDK network approximating the 0 more than the other two networks. In any case, the three negative values suggest that all three networks are more similar to a regular lattice than to a random graph.
With these measure on hand, we can conclude that while all three networks show properties of small-world networks, the one belonging to the professors of the DHDK course is the one in which the small-world phenomenon appears to be stronger, indicating a higher number of clusters and possibly a more robust resilience to perturbations.

6 Conclusion

In conclusion, our examination of the three co-authorship networks paint a few clear insights about the collaborative dynamics within the academic community of each course. The DHDK network, combining computer science and humanities, appears as an overall more cohese and connected network, displaying high small-world properties, and suggesting an efficient information flow among its components. Similar traits are found in CS and IT networks, as expected from these types of networks, albeit in quite smaller measures.
Overall we can conclude our research stating that the network of the DHDK authorship exhibits indeed a more cohesive community structure, in which the computational and humanistic realms come into a very closely knit and cohese structure capable of supporting itself quite strongly.

The results coming from the analysis of the centrality also seem to mirror the "hierarchy" of the course quite well, with degree director Silvio Peroni and ex-director Francesca Tomasi occupying the top position in many of the different measures for the DHDK course, reflecting their influence in connecting and perhaps in some cases even steering the academic contributions of the course. The same does not seem to happen for the CS and IT networks. The DHDK network emerges as a distinctive hub for promoting connectivity, cross-disciplinary dialogue, and innovation in the academic landscape.

7 Critique

Finally, we believe that our research provides a satisfactory answer to the research questions posed in the first sections of this report.
The study presented, however, quite a few problems during its devising and actuation phases. Firstly, in the construction of the networks, we decided to leave out from the graph different kinds of precious information, such as the more precise academic field of each teacher (which would have enabled a more meticulous study of the homophily). This decision was motivated by impossibility of coming up with a programmatical approach to solve the many inconsistencies and missing data found in the records available online in a short enough time. Additionally the choice of the "affiliation" attribute could lead to some inconsistencies due to its rather general an variable nature, as it is an attribute that could easily change over time.
Moreover, integrating our graph with completely new data would surely have benefited the preciseness of the measures. For example, adding metrics such as the h-index of each author would have allowed us to add some sort of empirical importance to each node, enabling more in-depth observations. Or perhaps, we could have included in the analysis the publication dates of each entry of our dataset. This omission limits our ability to grasp the temporal evolution of co-authorship networks. Understanding how collaborative patterns change over time is essential for discerning trends, identifying influential periods, and exploring the impact of academic events. Integrating the temporal dimension would have allowed us to investigate the longevity of interdisciplinary collaborations and distinguish between short-lived and enduring partnerships.