Proceedings of the
13th International Conference on Applied Informatics
Eger, Hungary, February 19–21, 2026
DOI: 10.17048/icai.2026.271
pp. 271–280

Local Scientific Collaboration Networks: A Case Study

Imre Varga*

Department of IT Systems and Networks,
Faculty of Informatics,
University of Debrecen
varga.imre@inf.unideb.hu

Abstract. This study analyzes local scientific collaboration through weighted co-authorship networks of two Hungarian university faculties using the domestic scientific bibliography (MTMT) data. Researchers are modeled as nodes and joint publications define weighted edges. Network metrics and a permutation-based statistical framework are applied to compare structures. While some global properties are similar, significant differences appear in degree and strength distributions, indicating distinct collaboration patterns independent of faculty size. The results highlight structural features that cannot be captured without a network-based approach. These insights are valuable for institutional strategy and decision-making in fostering collaboration.

Keywords: co-authorship, weighted network, MTMT, scientific collaboration

2020 *Mathematics Subject Classification:* Primary 05C82, 91D30, 05C90; Secondary 90C35

1. Introduction

Scientific collaboration networks have become a central object of study for understanding how knowledge is produced, disseminated, and validated within the scientific system. By modeling researchers and their joint publications as evolving graphs, publication networks make it possible to quantify patterns of cooperation, specialization, and interdisciplinary exchange at multiple scales. These networks

*The author thanks Bálint Kósa for the feasibility study.

capture not only who collaborates with whom, but also how scientific communities emerge, fragment, and reorganize over time in response to funding structures, institutional incentives, and emerging research topics [9, 10].

State-of-the-art research in this area combines large-scale bibliographic data with methods from network science, scientometrics, and computational social science. Recent advances include the identification of core–periphery structures, the detection of research communities and topical clusters, and the modeling of collaboration dynamics using temporal and multilayer networks. In particular, publication networks are increasingly used to study productivity, impact, and inequality in science, as well as to evaluate how collaborative structures influence innovation and the diffusion of ideas across disciplinary boundaries [5, 7].

Related works differ in many ways. They use different data sources e.g., Scopus [3], Web of Science [1], MedPub [10], arXiv [6], Google Scholar [4]. In these works, several problems are solved in different ways. For example, in some data sources, the authors cannot be identified unambiguously due to the different spellings or abbreviations of their names. Sometimes, each differently spelled name appearing in publications is considered a unique author, or authors with the same initials are considered identical authors [10]. We may face a similar problem with affiliations, if they are even specified in the databases. Some databases are not validated (e.g., Google Scholar), others cover only a narrow set of indexed journals (e.g., Scopus or WoS), but not all the publications of a given author.

In this work, we investigate the collaboration of faculty members of Hungarian universities using co-authorship networks of local scientific communities. Our goal is to quantify, understand, and visualize inter-personal and inter-department scientific collaborations through publications affiliated with the chosen universities. Earlier work has shown that node characterization based on the local graph environment may reveal structural properties that are not captured by purely global network measures [2].

2. Methodology

2.1. Data source

The present study builds on the publication data from the Hungarian Scientific Bibliography called MTMT [8], which provides a uniquely suitable data source for the analysis of scientific collaboration networks within the Hungarian research ecosystem. It is a national bibliography database for authentic recording and presentation of the results of domestic scientific research. MTMT records all scientific works, not just indexed articles. It has multiple levels of data validation: author upload and institutional review. Its use is mandatory, in a sense, so the majority of the Hungarian scientists have up-to-date MTMT accounts.

The MTMT MyCite2 API provides programmatic access to the database, enabling large-scale, automated retrieval of structured bibliographic and scientometric data. Through its RESTful interface, users can query publication records, author profiles, and citation links. The API facilitates reproducible data collection workflows by allowing parameterized queries, filtering by document types or publication years. Importantly, the standardized and curated nature of MTMT data, combined with the API's machine-readable output formats (e.g., JSON), significantly reduces the need for post hoc data cleaning and disambiguation.

This work focuses on the two organizing faculties of the *13th International Conference on Applied Informatics (ICAI2026)*, namely:

- Faculty of Informatics, University of Debrecen (UD),
- Faculty of Informatics, Eszterházy Károly Catholic University (EKCU).

They are faculties of prestigious eastern Hungarian universities. At the time of the study (December 2025), the MTMT profile identifiers of full-time lecturers/researchers at the above universities were collected and the publication records of these authors were downloaded and filtered by an in-house developed *Python* program.

The study only considered those publication types that represent significant scientific work and extensive publications:

- Book (e.g., Monography, Conference volume),
- Book portion (e.g., Book chapter, Essay),
- Journal article (e.g., Short, Summary, Multi-authored),
- Conference paper (e.g., proceedings paper).

Shorter and simpler publications (conference abstracts, reviews, critiques, theses, repertoires, etc.) were ignored.

The data gathering, the cleaning, the network formation, and its analysis are implemented by self-developed *Python* programs using the *NumPy* package. The visualization of different co-authorship networks is made using the *Cytoscape 3.10.1* software.

2.2. Network creation

From the downloaded publication records, the author IDs are extracted for each publication. These authors' lists of publications served as the basis of our local network definition. A $G = (V, E, W)$ weighted graph is defined for each organization, where a $v_i \in V$ node represents an active full-time researcher of the given faculty at the end of 2025. Nodes v_i and v_j are connected by an edge $e_{ij} \in E$, when there is at least one publication in the filtered data set co-authored by both v_i and v_j researchers. The scientific collaboration of two authors during their full career is characterized by the number of joint publications, which is represented by the $w_{ij} \in W$ weight of the e_{ij} edge. It is important to highlight that the graphs contain only the current employees (no retired, deceased or terminated employees), but all of their publications are considered, even if the author used to have a different affiliation earlier. The k_i degree of node i is the number of linked nodes. The s_i strength of node i is determined as the sum of w_{ij} weights of the connected edges, so $s_i = \sum_j w_{ij}$. This definition means that both the large average number of co-authors and the total number of co-authored publications increase this s_i collaboration strength of the author. The cardinality of the nodes is denoted by $N = |V|$, the cardinality of edges is denoted by $L = |E|$, and all w_{ij} weights are positive integers. The G graph is undirected (thus $w_{ij} = w_{ji}$) and loopless (so $e_{ii} \notin E$).

In some analyses, the departments of the authors are also considered, defining disjoint subsets of nodes, denoted as $V^d \subset V$. In this case, the set of edges can be divided into two disjoint subsets $E = E^{inter} \bigcup E^{intra}$, where E^{inter} means inter-department edges linking authors of different departments and E^{intra} means intra-department edges connecting authors from identical department.

2.3. Statistical network comparison

In $G_1 = (V_1, E_1, W_1)$ and $G_2 = (V_2, E_2, W_2)$ weighted networks of different sizes and topologies, we determine whether the distribution of a network quantity is truly different, or whether the difference can be explained purely by the network structure and sampling. Let $\gamma_i(v)$ denote a node metric for $v \in V_i$ node (or an edge feature for $e \in E_i$ edge). The observed differences between networks can be defined by the $\langle \gamma_i(v) \rangle$ averages of the given characteristics within the graphs as

$$\Delta_{obs} = \langle \gamma_1(v) \rangle - \langle \gamma_2(v) \rangle = \frac{1}{|V_1|} \sum_{v \in V_1} \gamma_1(v) - \frac{1}{|V_2|} \sum_{v \in V_2} \gamma_2(v).$$

Null hypothesis: the values of the given characteristics can be exchanged within the network so that the topology is preserved and the two networks come from the same generating mechanism. Our perturbation null model has the following steps:

1. We fix the structure of the G_1 and G_2 graphs.
2. We collect all the node (edge) characteristics:

$$\Gamma = \{\gamma_1(v) : v \in V_1\} \cup \{\gamma_2(v) : v \in V_2\}.$$

3. We randomly redistribute these values of characteristics among the nodes (edges), so that G_1 and G_2 each receive the same number of values as they originally had.

This is a conditional random sample, and we calculate this for each permutation: $\Delta^{(k)} = \left\langle \gamma_1^{(k)} \right\rangle - \left\langle \gamma_2^{(k)} \right\rangle$ This gives the null distribution: $\Delta^{null} = \{\Delta^{(1)}, \ldots, \Delta^{(M)}\}$. The statistical p-value can be defined as $p = P(\left|\Delta^{null}\right| \geq |\Delta_{obs}|)$ or empirically

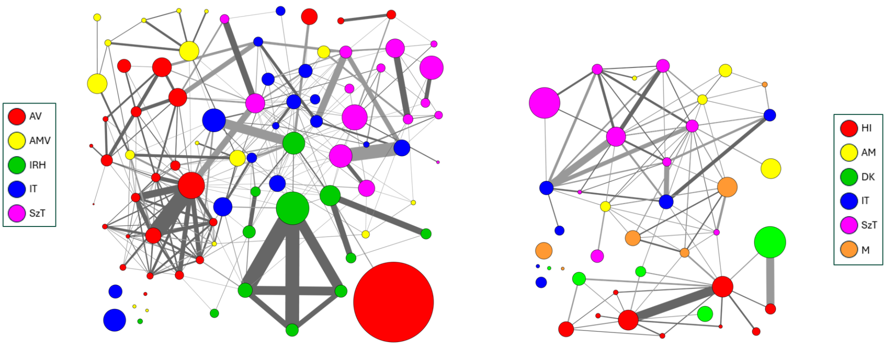
$$p \approx \frac{1}{M} \sum_{k=1}^{M} 1(\left|\Delta^k\right| \geq |\Delta_{obs}|).$$

If $p < 0.05$, then the difference cannot be explained solely by the structure of the network and the random characteristics distribution.

3. Results

First, a broad scope network analysis was performed, where not just the faculty members are considered, but all the co-authors of the faculty members, so the node set is extended. The network of the Faculty of Informatics, University of Debrecen (hereafter UD) contains 855 nodes (including the 84 faculty members) and 4605 links connecting them based on 5184 considered publications. The network of the Faculty of Informatics, Eszterházy Károly Catholic University (hereafter EKCU) contains 372 nodes (including the 40 faculty members) and 1065 links connecting them based on 1807 considered publications. However, the network of UD is much larger, network analysis shows that the $\langle k \rangle$ average degree of nodes representing the average number of co-authors of a researcher is not so different, for UD $\langle k \rangle = 22$, for EKCU $\langle k \rangle = 18$. It shows that the personal connections of the researchers are not purely determined by the size of an organization. Naturally, the degree distribution is not uniform; there are hubs in the networks. An interesting finding emerges from this study: the deans of both faculties have the most co-authors, despite the fact that their academic careers are far from the longest.

The remaining part of the investigation focuses on only the local co-authorship networks, covering only the researchers of the given faculty. This restriction causes smaller networks and easier visualization. The network of UD consists of $N = 84$ nodes and $L = 270$ links, as it is illustrated in Figure 1a, while the network of EKCU has $N = 40$ nodes and $L = 88$ links, see Figure 1b.



(a) UD **(b)** EKCU

Figure 1. Local co-authorship network of faculties. The area of nodes is proportional to the number of publications of the given author, and their colors represent the departments. The width of links is proportional to the edge weight, and different shades represent the intra- and inter-department connections.

As Figure 1 shows, there are tiny nodes of mainly young researchers with only a few publications, but there are big nodes as well, illustrating a large number of publications of productive researchers. An exceptionally large node is also visible in the UD network, representing a researcher frequently participating in nuclear physics projects (however, the faculty is related to computer science), where a large number of papers with many authors are common. (This author has 1822 publications, nevertheless the average number of considered publications for other UD scientists is just 51.) Wide links between two nodes show a lot of joint publications, illustrating strong collaborations. The maximum belongs to an author pair having 52 jointly written papers.

These local networks are not connected, since the majority of the nodes form a giant component, but there are isolated nodes without connections to others. It means that some of the researchers have never published together with others from the given faculty. The size of the isolated node is not always small, so we cannot assume these represent young scientists. One of the isolated nodes at UD has more than 140 publications affiliated with the given organization over decades, but none of them is co-authored by other faculty members, because this researcher has completely different research topics than others. An experienced researcher of the EKCU has 33 publications, but has not got links to other nodes, because he was recently hired on the faculty. The ratio of the giant component in both networks is close to 90%, however, the size of UD network is twice that of the EKCU.

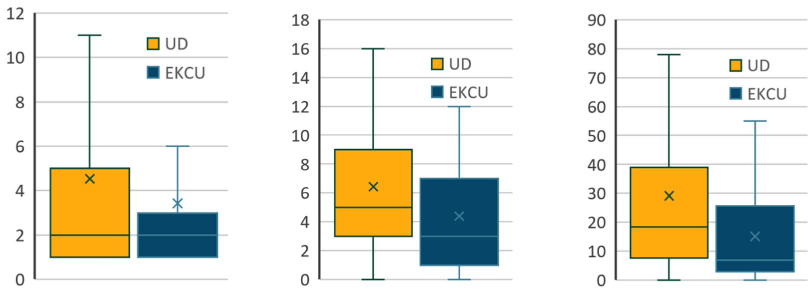
Further network characteristics are also calculated for both local collaboration networks in order to highlight differences or similarities between these graphs [12]. The features are summarized in the Table 1.

Table 1. The main network characteristics for the studied co-authorship networks.

Network metrics	**UD**	**EKCU**
Number of Components	7	6
Giant Component	92.9%	87.5%
Characteristic Path Length	2.658	2.582
Clustering Coefficient	0.419	0.513
Density	0.078	0.113
Average Degree	6.923	5.029
Minimum Degree	0	0
Maximum Degree	30	12
Network Diameter	6	5
Network Radius	3	3
Heterogeneity	0.725	0.696
Centralization	0.308	0.217

The average degree, the density, the centralization, or the average clustering coefficient are very different for these organizational networks. At the same time, the characteristic path length, the network radius, or the heterogeneity is almost the same. As one can see, from a given aspect, the networks are similar, but from other aspects, they are quite different. In order to determine whether these differences are caused by only the different size of the networks or these are structural differences, a statistical comparison of these weighted networks is needed.

The statistical comparison of weighted networks requires methods that account for the inherent dependence structure of relational data. To evaluate whether two networks differ in their weight- or structure-related properties, we adopt a permutation-based framework [11] that avoids the independence assumptions of classical parametric tests. Let G_1 and G_2 denote two weighted networks, and let $T(G)$ represent a network-level statistic of interest, such as the mean edge weight, mean node degree, or mean node strength. The observed difference between the networks is defined as $\Delta_{obs} = T(G1) - T(G_2)$. To assess statistical significance, we construct an empirical null distribution by pooling the relevant values (e.g., edge weights or node-level measures) from both networks and repeatedly permuting them while preserving the original sample sizes. For each permutation, the statistical difference is recomputed, yielding a distribution of differences expected under the null hypothesis that both networks arise from the same underlying distribution. The two-sided p-value is then estimated as the proportion of permuted differences whose absolute value exceeds that of the observed difference. This procedure is applied separately to each network metric, resulting in independent significance assessments for w_{ij} edge weights, k_i node degrees, and s_i node strengths. The box plot of these 3 characteristics is shown in Figure 2.

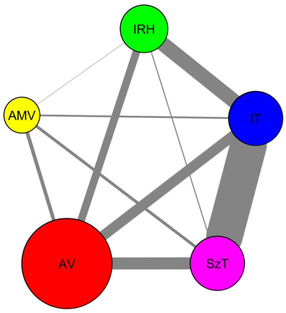


(a) Edge weights. **(b)** Node degrees. **(c)** Node strengths.

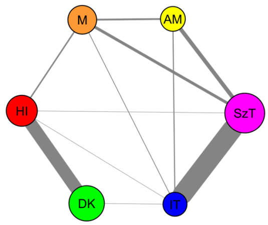
Figure 2. The box plot of the values of different network metrics for UD (orange) and EKCU (darkblue).

The box plots represent different statistical samples, but the statistical significance of the differences of metrics for these two compared weighted co-authorship networks can be captured by the p-values. This investigation revealed that link weights in these two networks there are no relevant differences ($p = 0.1915$). However, the collaboration networks of UD and EKCU are significantly different from the point of view of the node degrees (where $p = 0.0291$) and node strength (where $p = 0.0121$).

The researchers of a given department can be grouped together into one node representing the department within the faculty. In this way, one can focus on only the $e \in E^{inter}$ inter-department links. Results are illustrated in Figure 3.



(a) UD



(b) EKCU

Figure 3. Department-level co-authorship network of faculties. The area of nodes is proportional to the number of publications that belong to the given department. The width of links is proportional to the edge weight.

One of the most conspicuous differences between the two faculties is that, although the UD network is much larger than the EKCU network, the latter is still divided into more departments. This results in smaller departments with far fewer publications. One can see in Figure 3, that there are strongly and weakly connected departments in both faculties. This can be explained on the one hand by the different research areas of the departments, but we should also not ignore the fact that only a few research collaborations belonging to different departments can dominate the collaborations between departments that seem significant at first glance. In the latter, the restructure of the organization based on research interest could strongly influence inter-department collaborations.

4. Conclusions

This study set out to investigate the structure and characteristics of local scientific collaboration networks at the faculty level, with a particular focus on two Hungarian institutions (both faculties have the same research fields, namely computer science). By constructing weighted co-authorship networks from the Hungarian MTMT database, the research aimed to quantify interpersonal and interdepartmental collaboration patterns, and to reveal structural similarities and differences between the examined academic communities.

Methodologically, the work combined curated bibliographic data with tools from network science and statistical analysis. Co-authorship relations were modeled as weighted, undirected graphs, where edge weights captured the intensity of collaboration. In addition to standard network measures (e.g., degree, strength, clustering coefficient), a permutation-based statistical framework was applied to rigorously assess whether observed differences between networks could be attributed to structural effects or random variation. The analysis was performed at multiple levels, including extended collaboration networks, restricted local faculty networks, and aggregated department-level structures.

The results demonstrate that, despite differences in institutional size, certain global properties – such as characteristic path length and heterogeneity – are remarkably similar across the two faculties. At the same time, significant differences were identified in key structural metrics, including node degree and node strength distributions, indicating distinct collaboration patterns. The presence of hubs, the dominance of a giant component alongside isolated researchers, and the variability in inter-departmental connectivity all highlight the complex and heterogeneous nature of scientific collaboration at the local level. Furthermore, the analysis revealed that apparent differences in simple descriptive metrics (e.g., density or average degree) cannot be fully understood without statistical validation that accounts for network structure.

Importantly, these insights are inherently tied to the network-based approach. Many of the observed phenomena–such as the identification of central actors, the role of strong and weak ties, the structure of collaboration components, or the significance of inter-departmental links–cannot be captured by traditional, non-relational statistical methods. The network perspective enables the simultaneous consideration of individual behavior and system-level structure, thereby uncovering patterns that would otherwise remain hidden.

Beyond methodological contributions, the findings have practical relevance. Understanding the structure of local collaboration networks can support institutional strategy, for example, by identifying key researchers, potential collaboration gaps, or structurally isolated individuals and groups. At the department level, insights into interconnections may inform organizational design, interdisciplinary initiatives, and resource allocation. Consequently, such analyses can provide valuable input for local academic communities, faculty leadership, and higher-level decision-makers seeking to foster more effective and integrated research environments.

References

[1] T. Araújo, E. Fontainha: *The specific shapes of gender imbalance in scientific authorships: A network approach*, Journal of Informetrics 11.1 (2017), pp. 88–102, doi: 10.1016/j.joi.2016.11.002.

[2] C. Biró, G. Kusper: *Some k-hop based graph metrics and node ranking in wireless sensor networks*, Annales Mathematicae et Informaticae 50 (2019), pp. 19–37, doi: 10.33039/ami.2018.09.002.

[3] F. Catalá-López, A. Alonso-Arroyo, M. J. Page, L. Castelló-Cogollos, B. Hutton, M. Ridao, R. Tabarés-Seisdedos, R. Aleixandre-Benavent, D. Moher: *A cross-sectional analysis identified co-authorship networks and scientific collaboration on reporting guidelines for health research*, Journal of Clinical Epidemiology 157 (2023), pp. 22–34, doi: 10.1016/j.jclinepi.2023.02.001.

[4] Y. Chen, C. Ding, J. Hu, R. Chen, P. Hui, X. Fu: *Building and Analyzing a Global Co-Authorship Network Using Google Scholar Data*, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 1219–1224, doi: 10.1145/3041021.3053056.

[5] S. Drożdż, A. Kulig, J. Kwapień, A. Niewiarowski, M. Stanuszek: *Hierarchical organization of H. Eugene Stanley scientific collaboration community in weighted network representation*, Journal of Informetrics 11.4 (2017), pp. 1114–1127, doi: 10.1016/j.joi.2017.09.009.

[6] D. V. Filho, D. R. J. O'Neale: *Structure dynamics of evolving scientific networks*, arXiv, 2005.13251 (2020), pp. 1–16, doi: 10.48550/arXiv.2005.13251.

[7] M. J. Krawczyk, M. Libirt, K. Malarz: *Analysis of scientific cooperation at the international and intercontinental level*, Scientometrics 129 (2024), pp. 4983–5002, doi: 10.1007/s11192-024-05107-y.

[8] MTA: *MTMT*, 2026, url: https://www.mtmt.hu (visited on 01/10/2026).

[9] M. E. J. Newman: *Scientific collaboration networks. I. Network construction and fundamental results*, Phys. Rev. E 64 (1 2001), p. 016131, doi: 10.1103/PhysRevE.64.016131.

[10] M. E. J. Newman: *The structure of scientific collaboration networks*, Proceedings of the National Academy of Sciences 98.2 (2001), pp. 404–409, doi: 10.1073/pnas.98.2.404.

[11] S. L. Simpson, R. G. Lyday, S. Hayasaka, A. P. Marsh, P. J. Laurienti: *A permutation testing framework to compare groups of brain networks*, Frontiers in Computational Neuroscience 7 (2013), p. 171, doi: 10.3389/fncom.2013.00171.

[12] A. Zinilli: *Elements of Network Science*, Springer Nature Switzerland, 2025, doi: 10.1007/978-3-031-84712-7.