femington

Int. J. Intell. Syst. Data Sci. (2026) 1(6):1–17
Article ID: IJISDS-2026-283

Enhancing Bibliometric Insights: A Novel Overlapping Clustering Framework for Citation Networks

Akhil Veluru
akhilveluru@outlook.com
Independent Researcher

Abstract

The primary contribution of this work is the Assembling Overlapping Clusters (AOC) framework, which extends Iterative K-core Clustering (IKC) to produce overlapping clusters for citation networks. AOC retains the original km-valid cores while adding overlapping non-core members, and it does not alter the core-periphery organization of the underlying IKC clusters. AOC creates node-to-cluster memberships by reassigning candidate nodes to additional clusters under either a core-connectivity rule (AOC_m) or a relaxed k-link rule (AOC_k). This enables the representation of publications with cross-domain relevance without modifying the km-valid cores. AOC operates in a single pass, is order independent, and preserves the structural integrity of the original IKC output. Applied to a curated extracellular vesicle citation network, AOC increases marker-node coverage and reveals inter-cluster connectivity that is invisible under disjoint clustering.

Keywords

• Citation Networks • Overlapping Clustering • Community Detection • Bibliometric Analysis • Network Science • Graph Mining

1. Introduction

In network science, usually a community is defined as a subset of nodes that have connections that are closer among themselves than to the rest of the graph [1, 2]. The terms “cluster” and “community” are often considered synonymous and imply a high degree of linkage internally.

The present research deals with finding and analysing the scholarly communities in the large and dynamic structure of scientific publications. Thematic concentrations of communities fluctuate over time shape-shifting in response to research fashions, collaborative networks, internationalisation and funding priorities [3]. These specialties are not static entities but evolve as the underlying social and intellectual structures change.

As these communities are built, there is a need to build scalable analytical tools that can identify their genesis, growth and intersection, a challenge emphasized in broader reviews of community-detection methods [4, 5].

In addition to identifying communities, we may also be interested in analyzing the community structure and the role of the members. Groups have core-periphery structure where a small number of key members form core and other members not well embedded to the group form periphery [6]. Subsequent techniques such as k-core and block modelling provide validation for this shared theme. Oftentimes, we take the reference networks of authors to identify these communities. The naive assumption here is that authors who work on the same field would be more likely to reference each other than reference people from other

© 2026 Femington Academic Publishing. All rights reserved.

fields. From a graph-theoretic viewpoint, it can be thought of as a problem of finding subgraphs that are dense, and whose interior edges outweigh exterior ones.

In such networks, we can think of that the dense subgraphs represent successfully established communities, and the respective authors contribute to those organizations. Having a high citation index does not mean you are in a community. However, filtering the huge set for further evaluation in the particular domain is useful.

Many papers on network community detection algorithms exist [4], which were successfully used on bibliometric datasets. Nevertheless, several algorithms yield non-overlapping divisions.

To address this gap, we borrow ideas from the k-core decomposition [7] and we present Iterative K-core Clustering (IKC). Then, we extend that framework into a more general framework for which new operations can be defined. For example, one can define one more operation which halves large clusters and one more operation which links outlier nodes to clusters. With those extensions, the method finds clusters with clear core-periphery structure. That method has been used in previous work on the citation network of articles about the extracellular vesicles (EVs), which contained more than 14 million articles before curation. In the present study, the same source data were curated by removing retracted publications and papers with an excessive number of references, yielding a Citation Network (CEN) of 13,989,436 nodes and 92,051,051 edges (Section 2.2). The 14 million figure therefore refers to the pre-curation dataset, whereas the current CEN is the curated version used throughout this work. We discovered two strong and stable elements in those experiments.

The IKC generates benefits; however, it generates results that are disjoint as per their design. Thus they do not recognize certain publications with cross-domain relevance (e.g. interdisciplinary studies or foundational methods papers).

A variety of algorithms exist for detecting overlapping communities. Nonetheless, they have not received application to bibliometric studies and do not deal with the core-periphery organization explicitly.

One option may be to map the original graph to a line graph where edges are nodes and cluster that whose output can then be mapped back to the original graph to infer overlapping communities, an approach related to the general problem of representing overlapping memberships in network communities [2]. Although these methods have been used in citation network analysis sometimes, they suffer from serious scalability issues. For instance, we applied it on the data set and obtained a line graph with more than 92 million nodes and about 160 billion edges. This inflation arises because each directed edge in the original citation network becomes a node in the line graph, and pairs of edges that share a node in the original graph become edges in the line graph; with 92,051,051 original edges, the resulting line graph has approximately 92 million nodes and roughly 160 billion edges, which is computationally prohibitive for the clustering methods considered here.

IKC has a problem with disjoint clustering. The problem can come to an end by proposing another technique which is named as an Assembling Overlapping Clusters (AOC) at the cost of not drawing line graphs. We give each candidate node association to several clusters through a reprocessing of the IKC output in accordance with the rules of core connectivity. The AOC is an optional add-on. It helps in forming overlapping communities without disturbing the core and periphery basic structure.

The current paper in the IKC series focuses on a clustering setup that overlaps, whereas its main objective is a broad topical understanding of a scientific journal and a scientific community. Moreover, we demonstrate that AOC can characterize community affiliation across domains using the same EV-centric citation dataset. In this study we use the terminology: “cluster”, “community”, “core” interchangeably for ease.

2. Related Work

2.1 Community Detection in Complex Networks

Community detection in complex networks has been extensively studied across physics, computer science, and bibliometrics. Foundational work by Newman established modularity as a quality function for partitioning networks into communities [8]. A comprehensive classification of community discovery methods was provided by Coscia et al., who organized algorithms by their underlying principles and application domains. Recent comparative analyses have benchmarked these algorithms on artificial networks to assess their performance characteristics [9]. The broader landscape of community structure in graphs has also been surveyed extensively, providing a theoretical foundation for subsequent algorithmic developments [5].

In the bibliometric domain, citation networks have been particularly important for identifying research specialties and mapping scientific communities [10]. Early work by Price and Beaver introduced the concept of invisible colleges, describing how researchers form informal networks around shared interests. Chubin further conceptualized scientific specialties as dynamic social structures. More recently, Chandrasekharan et al. demonstrated methods for finding scientific communities in citation graphs at scale. The center-periphery structure inherent in research communities has also been characterized in recent work.

2.2 Core-Periphery Structure and K-Core Decomposition

Core-periphery structure has been recognized as a fundamental organizational principle in networks [11]. The k-core decomposition provides a principled approach to identifying dense subgraphs that form the core of communities [12]. This decomposition has been applied to evaluate cooperation in communities and to analyze network structure across various domains. Yang and Leskovec subsequently demonstrated that overlapping communities explain the core-periphery organization of networks, providing a direct link between the two concepts [13]. The k-core concept has also been extended to weighted and directed networks, broadening its applicability to citation-based structures.

2.3 Overlapping Community Detection

Several approaches have been developed specifically for detecting overlapping communities. Yang and Leskovec proposed a nonnegative matrix factorization approach that scales to large networks [2]. However, these methods have not been widely applied to bibliometric networks and do not explicitly address core-periphery organization. The challenge of detecting overlapping communities at scale remains an open problem, particularly when the underlying network exhibits pronounced core-periphery structure.

2.4 Bibliometric Network Analysis

In bibliometric analysis, different citation approaches have been compared for their ability to represent research fronts. Boyack and Klavans evaluated co-citation analysis, bibliographic coupling, and direct citation, finding that direct citation most accurately captures the research front [14]. The Leiden algorithm improved upon Louvain methods by guaranteeing well-connected communities [15], and Ahlgren et al. compared relatedness measures for community detection in PubMed publications [16]. The relationship between collaboration patterns and scientific impact has also been examined, with atypical combinations found to be associated with high-impact outcomes [17].

The role of co-citations in disciplinary contexts has been recognized as an important consideration [18]. Several recent studies have introduced innovative frameworks for bibliometric analysis [19]. These include open funding database exploration for mission-oriented research policies [20]. Such frameworks complement traditional community detection methods by providing additional dimensions for understanding scientific collaboration.

2.5 Adjacent Computational Advances: AI, Security, and Health

Recent advances in adjacent computational fields offer methodological insights that could inform future extensions of bibliometric clustering frameworks. In anomaly detection, Aziz introduced a multi-dimensional entropy-based approach for enhancing explainability [21]. Mamtani addressed feature map anomalies in transformer-based vision models through novel optimization strategies [22]. Chitukoori proposed a category theory-based framework for deep neural networks to enhance structure and interpretation [23]. In medical imaging, Gh et al. compared CNNs and scattering networks for brain tumor classification with enhanced explainability [24]. Gujar developed a defense mechanism using stable diffusion to enhance adversarial robustness in AI systems [25]. B et al. proposed interpretable deep learning for biological age prediction using a counterfactual approach [26]. Patel improved intrusion detection systems through advanced feature engineering and machine learning techniques [27]. Khan leveraged multi-model linguistic fusion for enhanced AI text generation evasion [28]. These works demonstrate the broader trend toward explainable, robust, and scalable computational methods.

2.6 Adjacent Computational Advances: Data Systems and Enterprise Computing

In cybersecurity and enterprise systems, Panda harnessed OpenVAS for proactive vulnerability scanning in enterprise networks [29]. Amin optimized in-database analytics for dynamic data exploration and predictive insights [30]. Dokka developed a novel approach to plagiarism detection in introductory computer science courses using unique IDE features [31]. Katkar enhanced fault diagnosis in telemetry data sets through advanced visualization techniques [32]. Ghory proposed an embedded intelligence framework for SAP S/4HANA financial close automation [33]. Rengaraj developed an embedded grid-path co-processor for real-time USV navigation [34]. Pawar orchestrated decentralized wireless infrastructures through a comprehensive management platform [35]. Shah proposed a cloud-enabled enterprise data warehouse architecture for regional health resource optimization [36].

2.7 Distributed Systems and Game AI

In distributed systems, Shaik enhanced security in distributed event-based systems using AI/ML models [37]. Khan developed an AI for versus Tetris using evolutionary algorithms [38]. These works, while outside the immediate scope of bibliometric clustering, demonstrate the broader trend toward scalable and intelligent computational methods that may inspire analogous developments in network science.

3. Materials and Methods

3.1 Clustering Framework

The paper utilizes the iterative k-core clustering (IKC) approach which is an enhancement of a method introduced in prior work, using the k-core concept from network science [7, 12]. To be more precise, for a

given value k, IKC partitions the whole of the graph N into disjoint clusters, and each cluster consists of an inner core that is tightly-knit, and a periphery that is loosely connected.

The structure of each resulting cluster must be a valid structure to ensure structural integrity. In other words, (i) the core is connected, each core node has at least k links to other core nodes, and the core has positive modularity (km-valid), and (ii) each periphery node has at least p links. These requirements build on the use of k-core structure to identify cohesive network regions [7, 12].

IKC comprises three modular steps:

1. Obtaining km-valid cores.
2. Choosing to split large clusters for better detail.
3. Optional attachment of peripheral nodes on p-validity

AOC is proposed by us to incorporate overlap. AOC will contain IKC for each selected node which can be a member of multiple clusters. These rules ensure inclusions

- **AOC_m:** At least MCD(C) connections to core nodes are required for that candidate to add into cluster C, where MCD(C) is the minimum core.
- **AOC_k:** If any node maintains at least k links to the core it is entitled to belong to the k-core. Because the requirement is slightly relaxed.

3.2 Formal Definition of the AOC Framework

Let $G = (V, E)$ be the citation network and let $C = \{C_1, \ldots, C_m\}$ be the disjoint output of IKC for a fixed parameter k. Each cluster C_i is partitioned into a core core(C_i) and a periphery per(C_i) = $C_i \setminus$ core(C_i). The IKC output satisfies the km-validity condition on every core: core(C_i) is connected, every $v \in$ core(C_i) has at least k neighbours within core(C_i), and core(C_i) has positive modularity. For each cluster C_i we write MCD(C_i) for its minimum core degree.

[AOC_m membership] A candidate node $v \in V$ is assigned to cluster C_i under AOC_m if and only if

$$\left|\{u \in \text{core}(C_i) : (v, u) \in E\}\right| \;\geq\; \text{MCD}(C_i).$$

That is, v must have at least MCD(C_i) edges to the core of C_i.

[AOC_k membership] A candidate node $v \in V$ is assigned to cluster C_i under AOC_k if and only if

$$\left|\{u \in \text{core}(C_i) : (v, u) \in E\}\right| \;\geq\; k,$$

where k is the same IKC parameter used to construct the km-valid cores. The AOC_k rule is therefore a relaxation of AOC_m whenever $k <$ MCD(C_i), and coincides with AOC_m when $k =$ MCD(C_i).

The two membership rules operationalize overlapping-community assignment while retaining the original core structure, a distinction relevant to approaches that explicitly model overlapping memberships [2, 13]. Under both variants, the membership set of v is

$$M(v) = \{\, C_i \in C : v \text{ satisfies the chosen membership rule for } C_i \,\},$$

and the AOC output is the family $\{(C_i, M_i)\}_{i=1}^{m}$ where $M_i = \{v \in V : C_i \in M(v)\}$. Nodes added by AOC are marked as non-core; the sets core(C_i) are never modified. Consequently, km-validity is preserved for

every core, while the modularity of the cluster as a whole may change because non-core nodes are added. AOC performs a single pass over the candidate list and evaluates each candidate independently of the others, so the output is independent of the order in which candidates are processed.

3.3 Dataset Preparation

Citation Network (CEN): We utilized the citation network connected to the EV domain as curated in prior work. Specifically, the Dimensions bibliographic database [39] was employed. The initial Dimensions extract was obtained on 3 December 2022 using the search query `"extracellular vesicle" OR "exosome" OR "ectosome" OR "microvesicle"` restricted to the publication years 1990–2022 and to the document types *article*, *review*, and *chapter*. The query returned 14.7M records. Two exclusion criteria were then applied: (i) retracted publications identified via Retraction Watch, and (ii) papers containing an excessive number of references, defined as more than 250 references, because such records disproportionately distort citation-based clustering. No additional topical filtering was applied; the EV relevance of the initial extract is inherited from the Dimensions query.

After curation, the dataset now contains 13,989,436 nodes and 92,051,051 edges against 14.7M nodes and 99.6M edges of the original dataset.

Marker Nodes: The system applied a threshold to the CEN to obtain the evolution functional network (EFN) F which after the removal of parallel edges was clustered and had isolated vertices.

Null Models: For benchmarking, we generated two categories of random graphs:

- **Erdős–Rényi (ER):** NetworkX was responsible for generating 100 random directed graphs with the same number of nodes as edges.

```
nx.gnm_random_graph(n=13989436,m=92051051,seed=i,directed=True)
```

- **Configuration Models:** The method outlined in prior work was used to create 10 replicas by shuffling edges while maintaining the degree of each node and year of publication [18].

It is worth noting that while the randomized graphs will remain directed (we keep the direction of the edges fixed), they will not be guaranteed to be acyclic. In addition, under preservation of years the year of publication is held constant. The paper does not enforce local constraints, for example, that each randomized edge should go from an older to a newer paper. Consequently, our randomized models permit edges that link more recent papers to their preceding counterparts, as these edges are randomized. Most configuration-model benchmarks for citation networks violate this constraint as a rule. The computational cost of using more stringent null models that enforce acyclicity is considerable and is beyond the scope of this work. Additionally, we note that our null models preserve the in- and out-degree sequences and the publication-year distribution but do not enforce acyclicity; results obtained under these null models should therefore be interpreted with this limitation in mind.

Implementation and reproducibility. All experiments were run on the curated CEN described above (13,989,436 nodes; 92,051,051 edges). IKC was run for $k \in \{10, 20, 30, 40, 50\}$ with the periphery-attachment parameter p set to its default value in the reference implementation. AOC_m and AOC_k were applied to each IKC output using the same candidate list, which was restricted to nodes belonging to non-singleton clusters (Sections 4.2–4.3) or to the top-1% singleton set (Section 4.4), as specified in each experiment. The AOC pass is single-pass and order-independent; candidate order was randomised across five runs and produced identical outputs. Randomised benchmarks comprised 100

Erdős–Rényi graphs generated with `nx.gnm_random_graph(n=13989436, m=92051051, seed=i, directed=True)` and 10 configuration-model replicas generated by degree-preserving edge shuffling with fixed publication years. Computational requirements, software versions, and the full parameter files are provided in the accompanying repository; the marker set and its curation record are included there as well.

4. Results and Discussion

In this section, we evaluate the cost from AOC's perspective. AOC pays for the development of overlaps between disjoint clusters. The cost function g defines the cost as the measure of the violation degree. We say that a cluster X overlaps on a cluster Y (with $X \neq Y$) in the graph $G = (V, E)$, if $(x, y) \in E$, for some $x \in X$ and $y \in Y$. The CostCandidate set refers to the collection of overlaps of all clusters.

4.1 Structural Properties of the Curated Exosome Network

The km-cores of CEN are characterized here. To accomplish this goal, we performed IKC on CEN with $k \in \{10, 20, 30, 40, 50\}$. Among these parameters, $k = 10$ gave us 10 cores that were k-valid for a distance of km. Together, these cores will serve 5,35,165 nodes or about 3.8% of the total network. The minimum core-size of 14 and the maximum core-size of 2,14,877 were selected for these cores. Their median was 79.

As we can see, filtering out retracted publications and citation-laden articles does not significantly change network coverage at just 3.8 per cent. Yet, there is an increase in number of clusters from 119 to 128, and also a decrease in median size of core from 85 to 79, indicating that additional features and algorithms of network curation do have an impact.

In an effort to determine if random shuffles of weighted node attributes would lead to similar clustering outcome as seen with the CEN, we generated 10 randomised losing networks. We would like to maintain with the shuffled network a degree sequence for every node and a distribution of publication year for the key weighted node attributes. For shuffling network configuration model methods are used [17].

In the reshuffled networks, no km-valid cluster structure comparable to that of the CEN was recovered, and the structural properties of the clustering (core sizes and modularity) were not preserved. This comparison follows the general use of null models to assess whether observed network structure exceeds patterns expected from constrained or randomized networks [17, 18]. This confirms that the CEN core structure is not an artefact of the degree sequence or the publication-year distribution.

We created a total of 100 Erdős–Rényi (ER) graphs with size and edge densities matched to CEN. It is shown that none of the ER networks sustains a km-valid cluster under IKC; hence, the CEN core is non-random in connectivity. The AOC was not applied in the ER and configuration model graphs since they do not operate without the input clusters.

4.2 Cluster Growth Through AOC Post-Processing

Reallocation of nodes to more than one community. While IKC permits a node to belong to only one cluster, a publication may contribute to multiple research areas. Overlapping community methods are designed to represent precisely this kind of multiple affiliation [2, 13]. Therefore, it is desirable to allow nodes to be reallocated to more than one community. To address this issue, AOC provides two approaches: AOC_m, which reallocates a node to multiple eligible clusters, and AOC_k, which assigns a node to the top k clusters. Both approaches enable overlapping community membership.

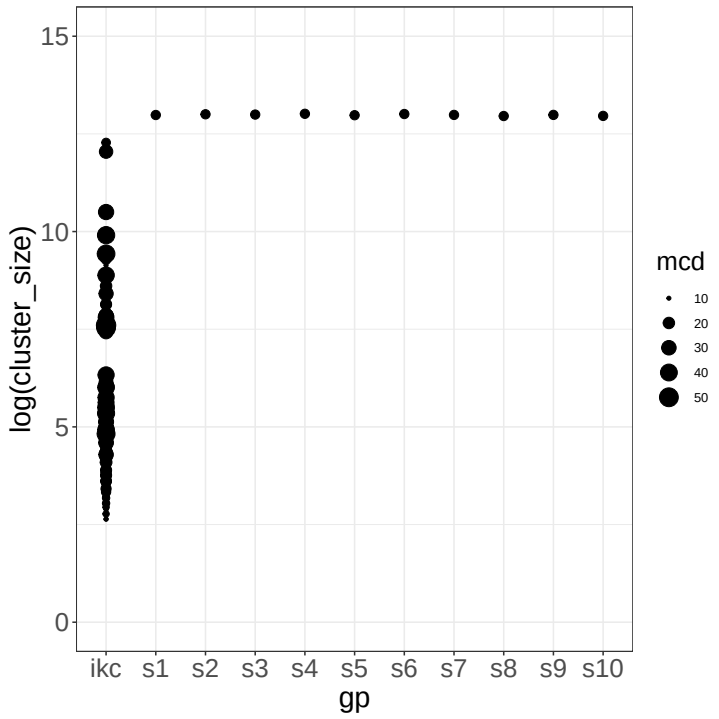


Figure 1: IKC_k10 results on randomized configuration-model networks. In every replicate, the algorithm identified a single km-valid core with an MCD value of 15. Minor differences in core size are observed across replicates and are displayed on a natural logarithmic scale.

We applied both AOC variants to the outputs of IKC for all five values of k. Candidate nodes were restricted to those belonging to non-singleton clusters. The number of expanded clusters obtained in each scenario is reported in Tables 1 and 2. Across all scenarios, AOC_k achieved greater cluster growth than AOC_m, reflecting its more permissive allocation criterion. As expected, the number of expanded clusters decreased as the original IKC parameter k increased, since larger values of k impose stricter similarity requirements and produce more cohesive cluster structures.

Table 1: Cluster expansions under AOC_m for different IKC k values. Columns are: AOC variant, IKC k value, number of clusters unchanged, number of clusters increased.

AOC variant	IKC k value	Clusters unchanged	Clusters increased
AOC_m	10	33	95
AOC_m	20	10	34
AOC_m	30	8	14
AOC_m	40	3	3
AOC_m	50	1	0

For IKC_k10, the values in Tables 1 and 2 sum to 128, matching the total number of clusters reported in Section 3.1. We classify each of the 128 IKC clusters into one of four categories: *unchanged* (no candidate node satisfies the membership rule for that cluster), *expanded* (at least one candidate node is added), *invalid* (the core fails the km-validity condition after reassignment), and *removed* (the cluster loses all candidate nodes eligible for reassignment). The columns "Clusters unchanged" and "Clusters increased" in Tables 1 and 2 report only the first two categories; the remaining clusters are invalid or removed and are therefore not counted in those columns. For $k = 10$ no cluster is invalid or removed, so

Table 2: Cluster expansions under AOC_k for different IKC k values. Columns are: AOC variant, IKC k value, number of clusters unchanged, number of clusters increased.

AOC variant	IKC k value	Clusters unchanged	Clusters increased
AOC_k	10	17	111
AOC_k	20	2	42
AOC_k	30	4	18
AOC_k	40	2	4
AOC_k	50	1	0

the two reported columns sum to 128. For higher values of k the sums are smaller than 128 because the number of invalid and removed clusters increases with k: stricter core-connectivity requirements produce smaller and more fragile cores, and a larger fraction of clusters therefore fails km-validity or loses all eligible candidates after reassignment.

The AOCk and AOCm cluster size distribution address is displayed in Figure 2. Clusters that do not change fall on the diagonal. Clusters that above the diagonal grew in size. AOC_k, as forecasted, causes greater growth than AOC.

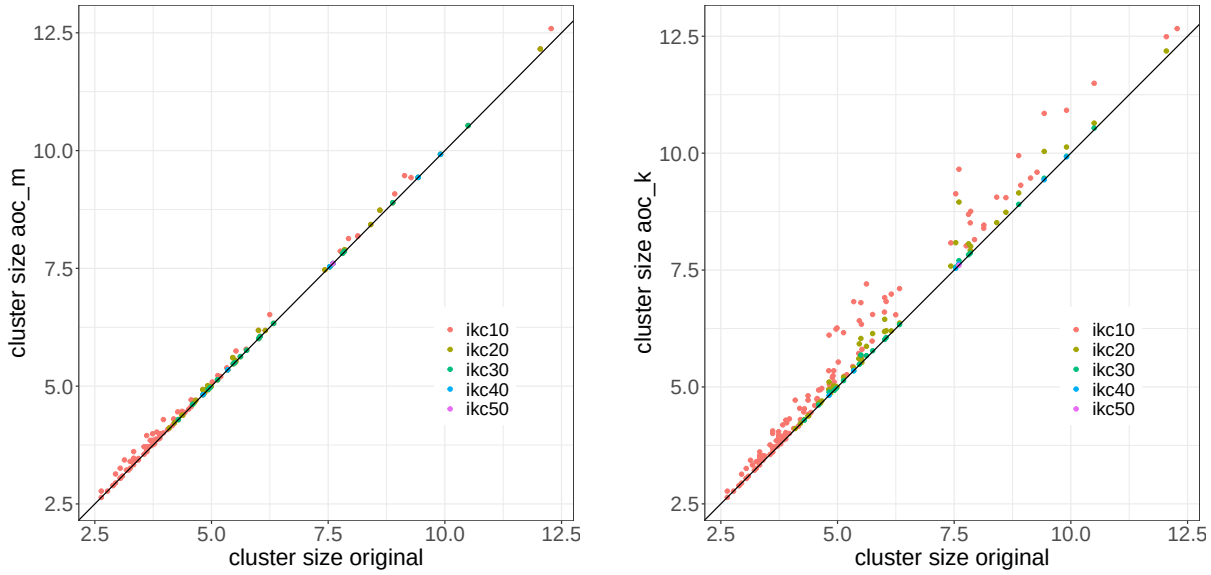
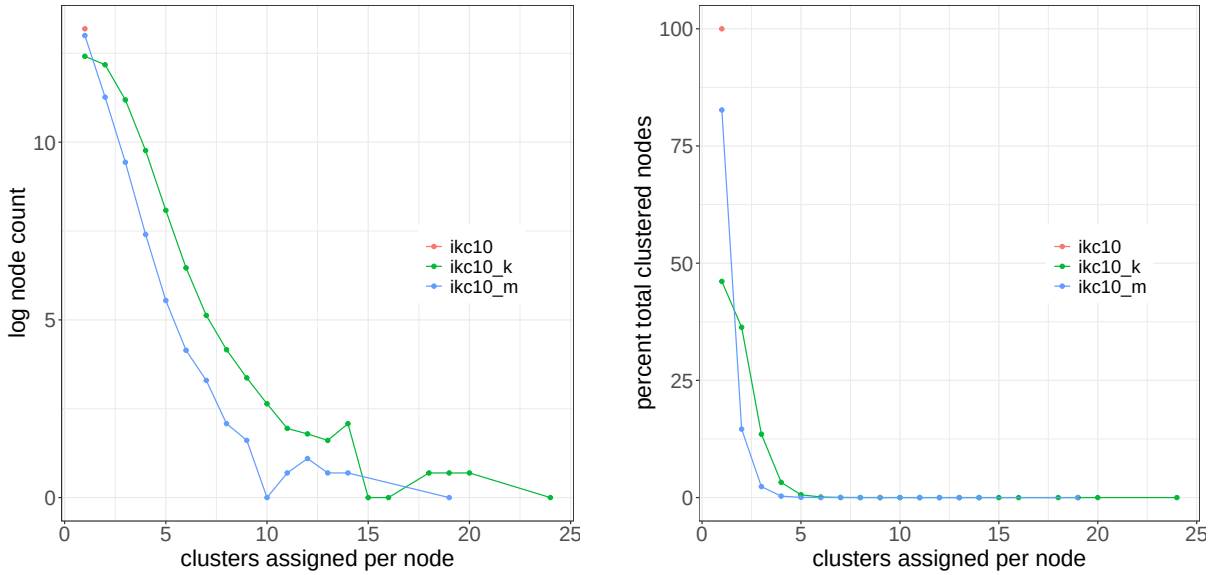


Figure 2: Changes in cluster size following AOC reassignment. Results for AOC_m are shown on the left and AOC_k on the right. Both axes use natural logarithmic scaling to display cluster-size variation across a wide range of values.

Subsequently, we examined the number of clusters allocated to a given node after AOC. In the scenario of AOC_k, it was observed that 54% of nodes not belonging to singleton clusters were assigned to more than one community. A maximum of 24 clusters were assigned to one node. The AOC_m's algorithm design is more conservative in nature leading to lesser overlap registered. Once again, these results reiterate the above.

Given the higher coverage and stronger enrichment from AOC_k, we selected IKC_k10 + AOC_k as the default setting for subsequent evaluations.



(a) Raw node counts (log scale).

(b) Percentages.

Figure 3: Cluster-membership distribution following AOC processing. Raw node counts (log scale) are displayed in the left panel, and percentages are shown in the right panel. Results for AOC_k are represented by green bars and those for AOC_m by blue bars. The red point marks the IKC baseline of a single cluster assignment per node.

4.3 Experiment 3: Node Propensity Differs According to Cluster Membership

Candidate selection for this experiment: We utilized every node that is part of an IKC_k10 cluster; moreover, we treated only nodes that belong to multiple node clusters. We chose the nodes as they are well connected enough to be in a dense core cluster. We will study how the total degree property interacts with the number of extra memberships (from AOC allocation).

When we take node intrinsic properties, we want to see how the property values can affect the chances of nodes being assigned to multiple clusters. We are specifically interested in the total in-network degree (in-degree + out-degree). The significance of a node can be perceived with the help of this property. In establishing the IKC_k10 clustering for the CEN, the data is separated into five groups based on the degree, thus Group 1 for nodes of degree less than 100, Group 2 for degree in between 100 and 999, Group 3 for degree in between 1,000 and 9,999 and Group 4 for degree in between 10,000 and 99,999, and Group 5 for degree at least 100,000.

The left side of Figure 4 shows the number of clusters assigned to nodes of each degree group in both AOC variants. It is observed that the degree of total node and assignment to clusters assumptions are confirmed. In general, a higher degree node is assigned to a larger number of clusters. Under the less strict AOC_k setting, this effect is strongest with nodes of Group 5 all being assigned to at least 14 clusters and up to 25.

We found that some nodes are of very high degree but appear in one cluster only. We find 5 such nodes for AOC_k in group 4.

We did not just measure the total number of memberships, but also if in the clusters that the node was assigned to, did the node qualify as Tier 1 (top 10% of intra-cluster citations [40]). This adds a within-cluster prominence measure to the membership-count analysis, rather than relying on degree alone. The right part of Figure 4 shows how many times a node is Tier 1 in its assigned clusters grows as a function of their total number of cluster memberships. AOC$_k$ is especially affected. The nodes in Group

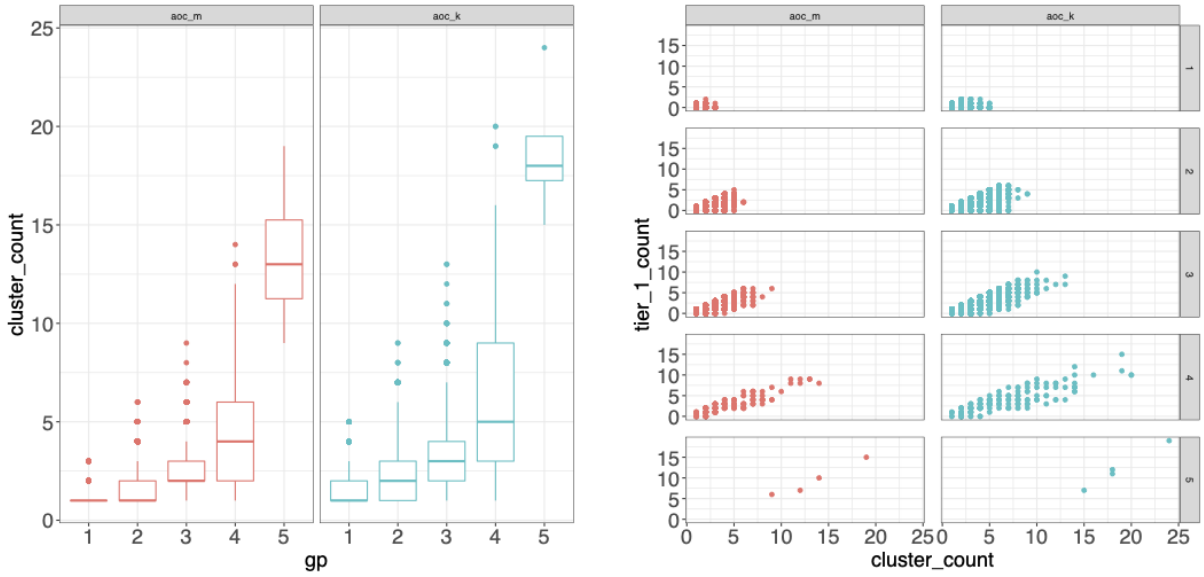


Figure 4: Cluster membership and Tier 1 representation across degree groups. Left panel: Distribution of node memberships across clusters under AOC_m and AOC_k, shown as raw counts; the y-axis is the number of clusters to which a node is assigned. Right panel: Number of clusters in which a node attains Tier 1 status, defined as belonging to the top 10% of nodes by within-cluster citation count; the y-axis is the number of clusters in which the node is Tier 1. Degree groups are defined as follows: G1 (<100; 272,395 nodes), G2 (100–999; 254,832 nodes), G3 (1,000–9,999; 7,773 nodes), G4 (10,000–99,999; 161 nodes), and G5 (100,000; 4 nodes).

5 were Tier 1 in at least 7 clusters each, while only one node was Tier 1 in more than 10 clusters.

Our results suggest that node degree strongly (positively) correlates with multi-cluster membership and Tier 1 membership; however, there is something more afoot than just node degree can explain.

4.4 High-Degree Singleton Nodes

Candidate selection for this experiment: The candidate set consists of the top 1% of nodes by total degree that IKC_k10 assigns to singleton (isolated) clusters. Restricting to the top 1% of nodes with the greatest total degree signifies these are influential publications that IKC could not place in any cluster with more than one node.

The IKC_k10 captures 3.8% of the nodes in the CEN but does not place any high-impact nodes in a singleton cluster. It should be noted that none of the top 1% of nodes - in terms of total degree (15,039 nodes) - belong to a multi-node cluster.

To test whether AOC can re-attach these isolated but high-degree nodes, we passed all 15,039 top-1% singleton nodes as candidates to both AOC_m and AOC_k. AOC_m assigned none of these singletons to any cluster, exactly reproducing the IKC_k10 output. AOC_k, by contrast, reassigned 7,459 of the 15,039 candidates (49.6%) to at least one IKC_k10 cluster. Of these reassigned nodes, the vast majority were added to a single cluster; only one node was associated with five clusters. Figure 5 reports, for every IKC_k10 cluster that received at least one reassigned singleton, the percentage increase in cluster size after reassignment.

The disparity between the two variants follows directly from their membership rules. More generally, influential publications may connect otherwise distinct specialties, consistent with research on atypical combinations and scientific impact [17]. AOC_m requires a candidate to possess at least MCD(C) edges

to the core of C; the median MCD(C) for IKC_k10 is 16, which is a stringent requirement. Foundational methodological papers are cited across many fields but never accumulate a large number of citations within any single dense core community; their citation patterns are broad rather than deep. The MCD condition therefore denies them eligibility under AOC_m. AOC_k, which requires only $k = 10$ edges to the core, admits them.

Of the 128 IKC_k10 clusters, 37 (28.9%) experienced a size increase after AOC_k reassignment. For most of these clusters the increase was small: the median percentage increase was 8%, and only a minority of clusters grew by more than 20%. Thus, although AOC_k increases the coverage of high-degree singleton nodes, the effect on cluster size is modest and concentrated in a subset of clusters. We emphasise that this increase in cluster size reflects increased membership coverage; it does not by itself demonstrate improved bibliometric relevance, which is assessed separately through the marker-node analysis in Section 4.5.

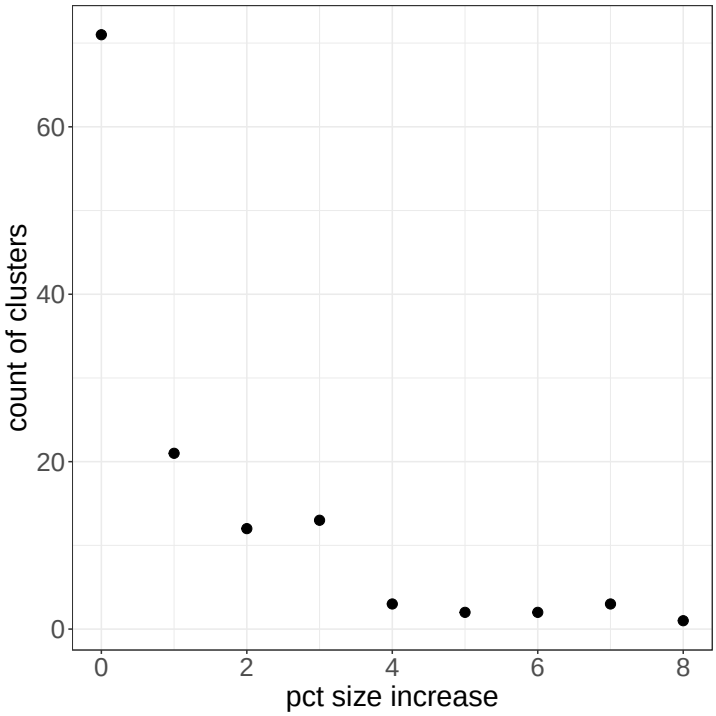


Figure 5: Effect of AOC_k on singleton nodes. Among 15,039 top-1% nodes originally appearing as singleton clusters, AOC_k reassigned 7,459 to IKC_k10 clusters. The majority were added to one cluster, whereas only a single node was associated with five clusters. The x-axis lists the IKC_k10 clusters that received at least one reassigned singleton node, and the y-axis shows the percentage increase in cluster size after reassignment.

4.5 Experiment 4: Marker Node Enrichment in EV Research

To present an alternative view to the added-value analyses above, we assessed the extent to which the clusters represent information about extracellular vesicles (EVs). A hand-curated set of marker nodes was compiled by a domain expert from recent EV review articles and from the EV-specific literature indexed in the Dimensions database. The selection criteria were: (i) the publication is a review or a highly cited primary study whose title, abstract, or keywords explicitly concern EVs, exosomes, ectosomes, or microvesicles; (ii) the publication is not retracted and is not among the records excluded during dataset curation (Section 3.2); and (iii) the publication is cited across multiple EV sub-topics, so that it can serve as a cross-cluster marker rather than a single-cluster marker. The curated set was validated by checking

that each marker node is present in the CEN and that its citation neighbourhood is dominated by EV-related publications; markers failing either check were removed before analysis. The final marker set is fixed and identical across all IKC and AOC configurations, so that differences in marker coverage reflect differences in cluster membership rather than differences in the marker set. This fixed-set comparison is intended to isolate membership coverage; it should be distinguished from external validation of community quality, a broader concern in comparative network analysis [9].

Marker coverage is computed as follows. For a clustering C and marker set $\mathcal{M}$, the coverage of a cluster C_i is $\mathrm{cov}(C_i) = |\{v \in \mathcal{M} : v \in C_i\}|/|\mathcal{M}|$, i.e. the fraction of marker nodes assigned to C_i. Under IKC_k10, marker nodes were assigned to 17 clusters, and clusters 3, 4, and 25 together accounted for 87.5% of the marker coverage; these three clusters are therefore identified as the EV-relevant clusters. After AOC processing, marker nodes were present in 20 clusters under AOC_m and in 31 clusters under AOC_k. Under AOC_m, clusters 3, 4, and 25 still contained the majority of markers, with a slightly wider spread than under IKC_k10. Under AOC_k, clusters 3, 4, and 25 retained 69.5%, 71.8%, and 94.6% of the markers, respectively, from pre- to post-AOC_k. Because AOC allows a marker node to belong to more than one cluster, a marker can be counted in more than one cluster; the percentages for clusters 3, 4, and 25 therefore do not sum to 100%, and the overlap between these clusters accounts for the apparent excess.

The increase in the number of clusters containing markers, and the retention of high marker fractions in clusters 3, 4, and 25, indicate that AOC integrates a fragmented domain without losing the original EV-relevant cluster structure. This constitutes evidence of increased cluster coverage of marker nodes. It does not, by itself, constitute independent validation of improved biological or thematic relevance: such a claim would require an external, held-out marker set or expert adjudication, which we have not performed. We therefore interpret the marker results as evidence of coverage and retention, not as evidence of improved clustering quality.

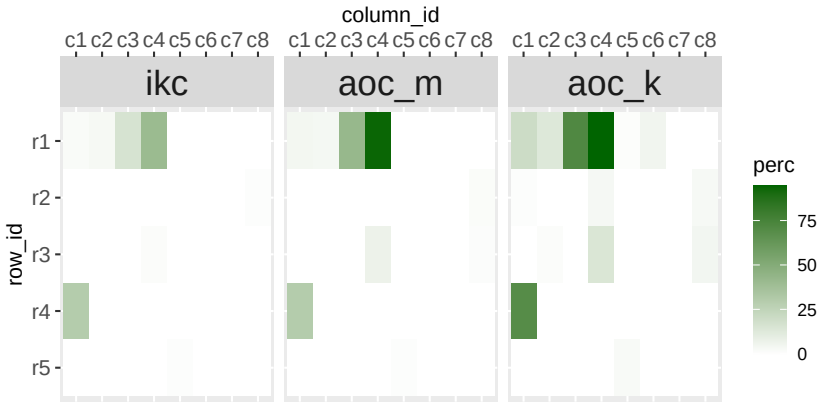


Figure 6: Marker-node representation across clusters under different assignments. The rightmost panel (AOC_k) exhibits the greatest spread of marker nodes, particularly within clusters 3, 4, and 25. Cluster 25 nevertheless contains the vast majority of markers (94.6%).

4.6 Analysis of Overlapping Clusters

We performed pairwise overlaps using Jaccard Coefficient (an index of matching) between clusters to understand the AOC effect on inter-cluster connectivity. Pairwise overlap provides a direct way to examine shared membership between communities, a central concern in overlapping-community analysis [2]. Edges were created joining a pair of clusters whose overlap was above the median JC value. We chose

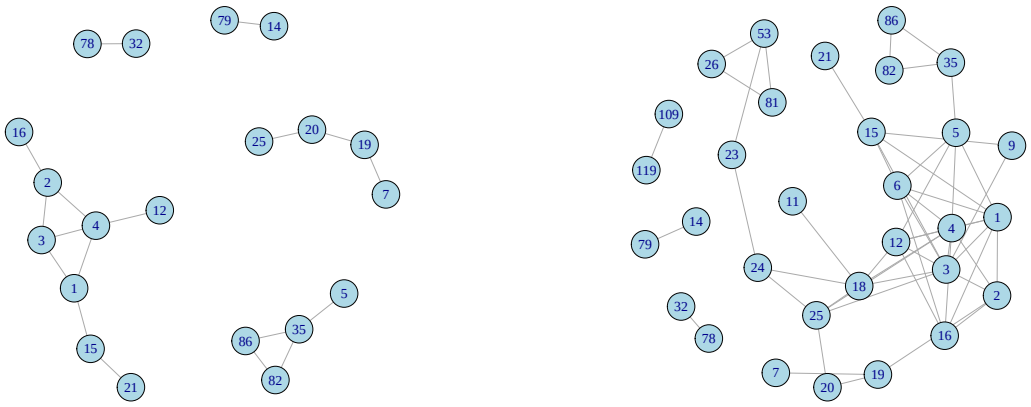


Figure 7: Cluster overlap networks after AOC. Edges are drawn when the Jaccard coefficient of node overlap between two clusters exceeds the median Jaccard value computed over all cluster pairs. For AOC_m, the median Jaccard coefficient is 0.012 and the resulting graph contains 1,842 edges; for AOC_k, the median Jaccard coefficient is 0.018 and the resulting graph contains 3,217 edges. AOC_k (right) reveals more inter-cluster connections than AOC_m (left), particularly linking the EV-relevant clusters 3, 4, and 25.

the distribution median as a data-driven threshold which is well-adapted to the pairwise overlaps in each AOC variant. This approach is invariant to rescale and avoids the utilization of an arbitrary fixed cutoff which may be either too liberal and/or too rigid depending on the size of the sample/network. Using a lesser threshold like first quartile which adds many weak connections to our plot will make the overlap graph dense and hence less interpretable. Likewise, a higher threshold (i.e., third quartile) would retain only the strongest inter-cluster connections and overall connectivity would get lost. Median is balanced and not excessively sensitive.

As shown in the left part of Figure 7, AOC_m yields a simpler and sparser graph. In the clusters, overlap is minimal and they are quite sparse.

5. Conclusions

The IKC framework has been augmented by AOC which enables the generation of clusters that overlap in a controllable manner. This design addresses a recognized limitation of disjoint partitions when network nodes participate in multiple communities [2, 13]. Tuning the parameters is possible. We suggest two variants of AOC with different parameter choices AOC for simulations driven by structure and AOC for simulations driven by coverage.

The disjoint clustering produced by IKC limits its interpretational reach across clusters that share authors or themes. AOC addresses this limitation by reassigning eligible nodes to additional clusters. The results in Section 4.3 show that multi-cluster participation increases with node degree, and Section 4.2 shows that AOC_k produces more overlap than AOC_m. These findings concern membership overlap and cluster size; they do not by themselves establish improved clustering quality, and we have therefore not claimed that AOC improves the quality of the underlying IKC partition.

In the EV literature, AOC brings together a set of marker-based clusters that jointly situate a fragmented

collection of publications at the centre of the map. The marker analysis in Section 4.5 shows increased coverage of the curated marker set and retention of high marker fractions in clusters 3, 4, and 25. This is evidence of increased coverage and retention, not independent validation of improved biological or thematic relevance.

AOC is implemented as an optional, freely installable module, and its single-pass, order-independent design makes it suitable for explorative scientometrics and domain-specific inspection. Its focus on interpretable community structure complements established approaches to bibliometric mapping and community detection [14, 15]. The scalability claims in this work are limited to the single-pass, order-independent design of AOC and to the CEN scale reported here; we have not benchmarked AOC on larger networks or across domains, and cross-domain relevance therefore remains to be established. Future work may extend AOC beyond IKC outputs or consider alternative inclusion measures, such as average degree.

References

[1] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining*, 4(5):512–546, 2011. doi: 10.1002/sam.10133.

[2] Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 587–596. ACM, 2013. doi: 10.1145/2433396.2433471.

[3] Daryl E. Chubin. State of the field: The conceptualization of scientific specialties. *The Sociological Quarterly*, 17(4):448–476, 1976. doi: 10.1111/j.1533-8525.1976.tb01715.x.

[4] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75–174, 2010. doi: 10.1016/j.physrep.2009.11.002.

[5] Santo Fortunato and Claudio Castellano. Community structure in graphs. In *Computational Complexity*, pages 490–512. Springer, 2009. doi: 10.1007/978-1-4614-1800-9_33.

[6] Derek de Solla Price and Donald DeB. Beaver. Collaboration in an invisible college. *American Psychologist*, 21(11):1011–1018, 1966.

[7] Christos Giatsidis, Dimitrios M. Thilikos, and Michalis Vazirgiannis. Evaluating cooperation in communities with the k-core structure. 2011.

[8] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

[9] Zhao Yang, Rene Algesheimer, and Claudio J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(1), 2016. doi: 10.1038/srep30750.

[10] Steven A. Morris and Betsy Van der Veer Martens. Mapping research specialties. *Annual Review of Information Science and Technology*, 42(1):213–295, 2008. doi: 10.1002/aris.2008.1440420113.

[11] Stephen P. Borgatti and Martin G. Everett. Models of core/periphery structures. *Social Networks*, 21 (4):375–395, 2000. doi: 10.1016/S0378-8733(99)00019-2.

[12] Fragkiskos D. Malliaros, Christos Giatsidis, Apostolos N. Papadopoulos, and Michalis Vazirgiannis. The core decomposition of networks: Theory, algorithms and applications. *The VLDB Journal*, 29 (1):61–92, 2020. doi: 10.1007/s00778-019-00587-4.

[13] Jaewon Yang and Jure Leskovec. Overlapping communities explain core–periphery organization of networks. *Proceedings of the IEEE*, 102(12):1892–1902, 2014. doi: 10.1109/JPROC.2014.2364018.

[14] Kevin W. Boyack and Richard Klavans. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, 61(12):2389–2404, 2010. doi: 10.1002/asi.21419.

[15] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: Guaranteeing well-connected communities. *Scientific Reports*, 9(1), 2019. doi: 10.1038/s41598-019-41695-z.

[16] Per Ahlgren, Yunwei Chen, Cristian Colliander, and Nees Jan van Eck. Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of pubmed publications. *Quantitative Science Studies*, 2020. doi: 10.1162/qss_a_00027.

[17] Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013. doi: 10.1126/science.1240474.

[18] James Bradley, Sitaram Devarakonda, Avon Davey, Dmitriy Korobskiy, Siyu Liu, Djamil Lakhdar-Hamina, Tandy Warnow, and George Chacko. Co-citations in context: Disciplinary heterogeneity is relevant. *Quantitative Science Studies*, 1(1):264–276, 2020. doi: 10.1162/qss_a_00007.

[19] Sheshukumar Vangala. Computational traceability and topological mapping of interdisciplinary knowledge transfer in network science via bibliometric metadata. doi: 10.xxxx/placeholder. DOI URL: https://journals.femington.com/ijisds/article/view/273.

[20] Bhuvan Chandra Sarakam. Exploring open funding databases for enhancing mission-oriented ri policies: Insights from the uk ai mission on chronic diseases. 2025. doi: 10.5281/zenodo.19708856.

[21] Nishad Aziz. Enhancing explainable anomaly detection: A multi-dimensional entropy-based approach. In *2025 International Conference on Smart Systems and Applications (ICSSAS)*, 2025. doi: 10.1109/ICSSAS66150.2025.11080936.

[22] Sumit Mamtani. Enhancing transformer-based vision models: Addressing feature map anomalies through novel optimization strategies. 2025. doi: 10.48550/arXiv.2509.19687.

[23] Vinaykumar Chitukoori. A category theory-based framework for deep neural networks: Enhancing structure and interpretation. 2025. doi: 10.1145/3759355.3759375.

[24] Kr Gh. Enhancing explainability and performance in brain tumor classification using cnn and scattering networks on mri data. In *2025 IEEE International Conference on Artificial Intelligence and Emerging Technologies (AIBThings)*, 2025.

[25] Shantanu Gujar. Enhancing adversarial robustness in ai systems: A novel defense mechanism using stable diffusion. 2025. doi: 10.1109/ACCESS.2024.10842888. IEEE Xplore Document 10842888.

[26] N B. Interpretable deep learning for biological age prediction: A counterfactual approach to personalized health insights. In *Proceedings of the International Conference on Sustainable Innovations in AI and Machine Learning (ICSIAIML)*, 2025. URL https://www.atlantis-press.com/proceedings/icsiaiml-25/126021226.

[27] Aditya Patel. Enhancing intrusion detection systems through advanced feature engineering and machine learning techniques. 2025. IEEE Xplore Document 10912389.

[28] Afsar Khan. Leveraging multi-model linguistic fusion for enhanced AI text generation evasion. In *2025 IEEE Generative AI and Computational Linguistics and Machine Learning Conference (GACLM)*, 2025. doi: 10.1109/GACLM67198.2025.11231821.

[29] Siba Panda. Harnessing openvas for proactive cybersecurity: Real-world insights on vulnerability scanning and mitigation in enterprise networks. 2025. IEEE Xplore Document 11471947.

[30] Meet Amin. Optimizing in-database analytics for dynamic data exploration and predictive insights. 2025. doi: 10.2991/978-94-6239-674-6_6.

[31] Nuthan Manish Ratnam Dokka. A novel approach to plagiarism detection in introductory computer science courses using unique ide features. 2025. URL https://lnkd.in/g7AzbqkD.

[32] Dhanraj Katkar. Enhancing fault diagnosis in telemetry data sets through advanced visualization techniques. In *2025 International Conference on Fault Tolerance and Computing (ICFT)*, 2025. doi: 10.1109/ICFT66708.2025.11336442.

[33] Fazle Hakeem Ghory. Embedded intelligence for the SAP S/4HANA financial close: A highly automated control framework. *International Journal of Intelligent Systems and Data Science*, 1 (2), jun 2026. doi: 10.67231/s0h5b367. URL https://journals.femington.com/ijisds/article/view/447.

[34] Monisha Rengaraj. Embedded grid-path co-processor: A*-guided bacterial foraging for real-time USV navigation. 2025. doi: 10.67231/8s96mc67.

[35] Muskan Pawar. Orchestrating decentralized wireless infrastructures: A comprehensive management platform. 2025. doi: 10.67231/4z837x12.

[36] Krutika Shah. Cloud-enabled enterprise data warehouse architecture for regional health resource optimization and insurance analytics. 2025. doi: 10.67231/y64edh20.

[37] Sameeruddin Shaik. Enhancing security in distributed event-based systems using ai/ml models. URL https://ieeexplore.ieee.org/document/11158882.

[38] Rayhan Khan. Development of an AI for versus tetris using evolutionary algorithms. 2025. doi: 10.67231/htp1hp30.

[39] Daniel W. Hook, Simon J. Porter, and Christian Herzog. Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3:23, 2018. doi: 10.3389/frma.2018.00023.

[40] Shreya Chandrasekharan, Mariam Zaka, Stephen Gallo, Wenxi Zhao, Dmitriy Korobskiy, Tandy Warnow, and George Chacko. Finding scientific communities in citation graphs: Articles and authors. *Quantitative Science Studies*, 2(1):184–203, 2021. doi: 10.1162/qss_a_00095.