computers

MDPI

Article

A Reproducible Computational Pipeline for Cross-Database Scientometric Network Construction: Architecture, Algorithms, and Structural Validation

Denny Moreno-Castro [1,2], Omar Orlando Franco-Arias [1], Cícero Pimenteira [2], Nicolás Márquez [3,*] and Cristian Vidal-Silva [4,*]

[1] Facultad de Ciencias e Ingeniería, Universidad Estatal de Milagro (UNEMI), Milagro 091050, Ecuador; dmorenoc@unemi.edu.ec (D.M.-C.); ofrancoa@unemi.edu.ec (O.O.F.-A.)
[2] Programa de Pós-Graduação em Ciência, Tecnologia e Inovação Agropecuária (PPGCTIA), Universidade Federal Rural do Rio de Janeiro (UFRRJ), Seropédica 23890-000, Rio de Janeiro, Brazil; cicero@ufrrj.br
[3] Escuela de Ingeniería Comercial, Facultad de Economía y Negocios, Universidad Santo Tomás, Talca 3460000, Chile
[4] Facultad de Ingeniería y Negocios, Universidad de Las Américas, Manuel Montt 948, Providencia, Santiago 7500975, Chile
* Correspondence: nmarquez4@santotomas.cl (N.M.); cristian.vidal.silva@edu.udla.cl (C.V.-S.)

Abstract

The rapid expansion of scientific publications indexed in multiple bibliographic databases has created new computational challenges for large-scale scientometric analysis. Differences in metadata schemas, identifier structures, and export formats across indexing systems such as Web of Science and Scopus introduce inconsistencies that may distort network-based bibliometric analyses. These issues affect duplicate detection, node identification, and network topology construction. This study proposes a reproducible computational pipeline for cross-database scientometric network construction. The framework formalizes the preprocessing workflow into explicit computational modules, including metadata harmonization, deterministic duplicate detection, sparse graph construction, normalization, and structural diagnostics. The proposed architecture separates preprocessing stages into reproducible algorithmic components, enabling transparent evaluation of methodological assumptions. Empirical evaluation using an interdisciplinary dataset of 317 publications (1990–2023) demonstrate that deterministic preprocessing significantly improves network stability and preserves clustering structure. Structural diagnostics based on modularity, Herfindahl–Hirschman Index, Shannon entropy, and Gini coefficient provide multi-dimensional evaluation of network topology. Scalability experiments confirm near-linear computational growth under sparse graph construction. The principal contribution of this work lies in the formalization of a transparent and extensible computational architecture for reproducible scientometric analysis. The proposed pipeline supports reliable cross-database integration and enables scalable knowledge-mapping applications in interdisciplinary research domains.

Keywords: scientometric computing; bibliometric network modeling; metadata harmonization; duplicate detection algorithms; reproducible research pipelines; graph-based knowledge mapping; computational scientometrics

Check for updates

Academic Editor: Paolo Bellavista

Received: 11 March 2026
Revised: 18 March 2026
Accepted: 26 March 2026
Published: 31 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid growth of indexed scientific production has transformed bibliometric analysis into a computationally intensive field that relies on large-scale metadata processing and network modeling [1–6]. Major bibliographic databases such as Web of Science (WoS) and Scopus provide structured metadata that enables quantitative analysis of scientific knowledge through citation networks, co-authorship graphs, and keyword co-occurrence structures [7,8].

Network-based bibliometric approaches are widely used to map research domains, identify emerging scientific topics, and analyze collaboration patterns across disciplines [1–4,7,9–11]. These approaches typically represent documents, authors, or keywords as nodes connected through citation or co-occurrence relationships, allowing the application of graph-theoretical methods to study the structural organization of scientific knowledge [12–15].

Despite their widespread adoption, bibliometric analyses increasingly face methodological challenges when integrating data from multiple indexing platforms. Differences in metadata schemas, identifier coverage, export formats, and classification systems introduce structural heterogeneity that can affect data consistency and downstream network analysis [16–18]. To illustrate these structural differences, Table 1 summarizes key characteristics of major bibliographic databases. These discrepancies complicate cross-database integration and may lead to issues such as duplicate records, fragmented author identities, and inconsistent keyword representations, ultimately impacting the reliability of scientometric results.

Table 1. Structural differences between major bibliographic databases.

Feature	Web of Science	Scopus
Author identifiers	Partial coverage	Broad coverage
Citation indexing	Highly curated	Broader journal coverage
Export formats	Structured field tags	Mixed metadata formats
Document classification	Detailed categories	Broader subject areas

Source: Authors' own elaboration.

Several software tools support bibliometric network analysis, including VOSviewer (version 1.6.19 or higher) and bibliometrix (version 3.0 or higher) [1,7,19]. These tools provide advanced visualization and clustering techniques; however, many preprocessing operations—such as metadata harmonization and duplicate resolution—are embedded implicitly within tool-specific workflows, limiting transparency and reproducibility [20,21]. When datasets exported from heterogeneous sources such as WoS and Scopus are combined, differences in metadata structures and identifier conventions may generate duplicate records, fragmented author representations, and inconsistent keyword groupings [16,17,22,23]. These inconsistencies can propagate into network construction and significantly affect structural indicators such as density, centrality distributions, and community structure [8,24,25].

Figure 1 illustrates the typical workflow of cross-database bibliometric network construction and highlights the preprocessing stages where structural distortions may arise. These issues are not only related to data integration but also to the lack of explicit, reproducible preprocessing procedures capable of isolating and evaluating the impact of each transformation step [18,26–28].

To address these challenges, this study proposes a reproducible computational pipeline for cross-database scientometric network construction. The objective of this work is to design and evaluate a modular preprocessing framework that enables consistent integration

of heterogeneous bibliographic data while preserving the structural properties of the resulting networks.

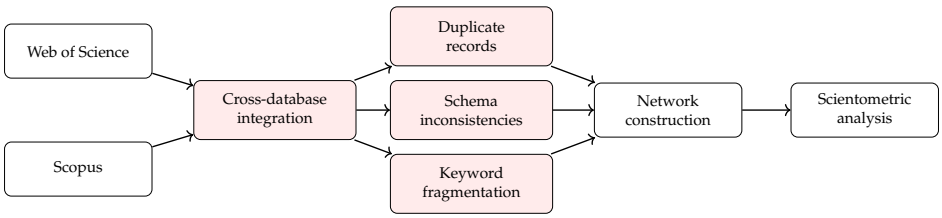


Figure 1. Sources of structural distortion in cross-database bibliometric network construction.

The proposed approach explicitly formalizes key preprocessing stages, including metadata harmonization, deterministic duplicate detection, and structured data preparation. Unlike many existing tools that emphasize visualization or exploratory analysis, this pipeline provides a transparent and reproducible workflow that allows the impact of each transformation step to be systematically analyzed.

From a broader perspective, the increasing reliance on multi-database scientometric analysis reflects a shift toward data-intensive research evaluation practices [21,29]. However, this shift also introduces methodological challenges related to data consistency, reproducibility, and interpretability [18]. Addressing these challenges requires not only computational efficiency but also methodological transparency, ensuring that preprocessing decisions can be systematically examined and validated. The main contributions of this work are summarized in Table 2.

Table 2. Main contributions of the proposed framework.

Contribution	Description
Computational pipeline	Modular preprocessing architecture
Duplicate detection	Deterministic record linkage algorithm
Network diagnostics	Structural concentration metrics
Scalability analysis	Evaluation under dataset expansion

Source: Authors' own elaboration.

This study is structured as an empirical research article, in which the proposed computational pipeline is formally defined and experimentally evaluated using real-world bibliographic data.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents the experimental results. Section 5 discusses the findings. Section 6 concludes the study.

2. Related Work

The rapid expansion of scientific publications has stimulated the development of computational techniques for analyzing large-scale bibliometric datasets [30]. Modern scientometric studies increasingly rely on network-based models that represent documents, authors, or keywords as nodes connected through citation or co-occurrence relationships [1,7]. These representations enable the application of graph-theoretical methods to analyze the structural organization of scientific knowledge and to identify thematic clusters within research domains [31].

Several software environments have been developed to support bibliometric and scientometric analysis. Tools such as VOSviewer and bibliometrix provide advanced functionality for network visualization, clustering, and science mapping [1,3,7]. These systems are widely used for exploring citation structures, co-authorship networks, and

keyword co-occurrence patterns across scientific domains. However, most existing tools assume internally consistent metadata and focus primarily on visualization and exploratory analysis rather than on reproducible preprocessing pipelines [5].

A key limitation of current bibliometric software environments lies in the implicit handling of preprocessing operations. Tasks such as metadata harmonization, duplicate resolution, and schema normalization are typically embedded within tool-specific workflows, limiting transparency and reproducibility when integrating heterogeneous bibliographic datasets [32,33]. As a consequence, the impact of data integration decisions on network structure is rarely evaluated in a systematic and reproducible manner.

Table 3 highlights that, while existing tools provide strong analytical and visualization capabilities, they offer limited support for explicit and reproducible preprocessing workflows. This limitation is particularly critical in cross-database scientometric studies, where inconsistencies in metadata schemas, identifier conventions, and export formats may introduce duplicate records, fragmented author identities, and distorted network structures.

Table 3. Comparison of major bibliometric analysis tools.

Tool	Main Focus	Network Analysis	Visualization	Reproducible Preprocessing
VOSviewer	Science mapping	Yes	Advanced	Limited
bibliometrix	Bibliometric statistics	Yes	Moderate	Partial
CiteSpace	Citation burst detection	Yes	Advanced	Limited
Proposed framework	Cross-database integration	Yes	Modular	Explicit pipeline

Source: Authors' own elaboration.

Graph theory provides a mathematical foundation for analyzing structural properties of bibliometric networks, including connectivity patterns, centrality distributions, and community structures [8]. Empirical studies have consistently observed heterogeneous degree distributions and modular organization in large-scale scientific networks, reflecting thematic segmentation and disciplinary subfields [8,13]. These characteristics support the use of community detection and structural diagnostics in scientometric analysis.

Despite their widespread adoption, existing bibliometric tools generally assume internally consistent metadata inputs. However, bibliographic databases often exhibit structural heterogeneity in metadata schemas and identifier systems [16]. When datasets from multiple indexing platforms are combined, inconsistencies in author names, journal titles, and citation formats may produce duplicate records or fragmented citation structures. These issues can propagate into network construction and affect structural metrics such as centrality, clustering, and modularity, ultimately leading to unreliable interpretations of scientific domains.

Recent studies have emphasized the importance of reproducible computational workflows in bibliometric analysis [21,29,34]. Explicit preprocessing pipelines enable researchers to systematically evaluate the impact of data integration decisions on network topology and clustering outcomes. In this context, there remains a clear gap in the literature regarding the formalization of preprocessing as a transparent and modular computational process.

This study addresses this gap by proposing a reproducible computational pipeline that explicitly separates metadata harmonization, duplicate detection, network construction, and structural diagnostics into independent and evaluable components.

3. Methodology

This section describes the methodology underlying the proposed computational pipeline for cross-database scientometric network construction. The approach is struc-

tured as a sequence of processing stages, including metadata harmonization, duplicate detection, network construction, and structural analysis.

From a network-analysis perspective, bibliometric workflows require careful preprocessing because structural inconsistencies in metadata may propagate into graph topology, centrality estimation, and clustering results [7,8]. Accordingly, the methodology isolates operations that are often embedded implicitly within bibliometric software environments, particularly schema alignment, record linkage, and weighted network generation [7,16].

3.1. Metadata Harmonization

The methodology begins with metadata harmonization, a preprocessing stage aimed at resolving schema heterogeneity across bibliographic databases such as Web of Science and Scopus. Cross-database integration requires aligning heterogeneous metadata fields including author identifiers, publication titles, keyword descriptors, and document classification categories [35,36].

Metadata harmonization is a critical stage in large-scale bibliometric analysis because inconsistencies in schema structure and field formatting may lead to fragmented author identities, duplicated records, and inconsistent keyword vocabularies [29,37]. Differences in export formats and metadata completeness across indexing platforms may introduce systematic biases when constructing scientometric networks. To address these issues, the methodology performs explicit normalization of metadata attributes before network construction.

The pipeline begins with metadata ingestion from independent bibliographic databases. Let D^{WoS} and D^{Scopus} denote the raw metadata collections exported from Web of Science and Scopus, respectively. Because these sources use different field names, identifier conventions, and export structures, schema harmonization is required [16]. Formally, a harmonization function H is defined to transform heterogeneous metadata into a unified representation:

$$D^{\mathrm{harm}} = H\left(D^{\mathrm{WoS}} \cup D^{\mathrm{Scopus}}\right) \tag{1}$$

This transformation standardizes core metadata fields including DOI, author names, journal titles, publication year, document type, and keywords. Consistent with prior work, string normalization is applied to reduce lexical fragmentation arising from punctuation, capitalization, and variant abbreviations [1,2]. Table 4 summarizes the harmonization rules.

Table 4. Metadata harmonization rules implemented in the proposed methodology.

Field	Harmonization Rule
DOI	Lowercasing, trimming, prefix normalization, null-value validation
Authors	Lowercasing, punctuation removal, whitespace normalization
Journal title	Standardization of abbreviations and typographic variants
Keywords	Tokenization, stopword filtering, lowercasing
Document type	Mapping to a unified categorical encoding
Publication year	Numeric validation and missing-value consistency checks

Source: Authors' own elaboration.

3.2. Deterministic Duplicate Detection

Following metadata harmonization, the second stage of the methodology addresses duplicate detection. Duplicate records are resolved through a deterministic two-stage procedure, as the same publication may appear in multiple indexing systems with partially inconsistent metadata representations [19,22].

In the first stage, exact matching is performed using persistent identifiers such as DOI. When available, DOI matching provides a reliable mechanism for bibliographic record linkage because it represents a globally unique identifier assigned to scholarly

publications [25]. However, DOI metadata may be missing or inconsistently formatted in some records.

When DOI information is unavailable or incomplete, a composite similarity score is computed by combining title similarity, author overlap, and publication-year agreement. Let S denote the similarity score between two records r_i and r_j:

$$S(r_i, r_j) = \alpha S_{\text{title}} + \beta S_{\text{authors}} + \gamma S_{\text{year}} \tag{2}$$

where S_{title} measures normalized title similarity, S_{authors} captures author-list overlap, and S_{year} represents publication-year agreement [23,38]. The coefficients satisfy:

$$\alpha + \beta + \gamma = 1 \tag{3}$$

A pair of records is classified as duplicate when:

$$S(r_i, r_j) \geq \tau \tag{4}$$

where τ denotes a predefined similarity threshold. This rule-based formulation emphasizes transparency and reproducibility.

Algorithm 1 summarizes the duplicate detection procedure.

Algorithm 1 Deterministic duplicate detection

Require: Harmonized metadata set D^{harm}, similarity threshold τ
Ensure: Deduplicated metadata set D^*
Initialize $D^* \leftarrow \emptyset$
for each record $r_i \in D^{\text{harm}}$ **do**
  mark $\leftarrow$ false
  for each record $r_j \in D^*$ **do**
    if DOI(r_i) is not null and DOI(r_i) = DOI(r_j) **then**
      mark $\leftarrow$ true
    else
      Compute $S(r_i, r_j)$
      if $S(r_i, r_j) \geq \tau$ **then**
        mark $\leftarrow$ true
      end if
    end if
  end for
  if mark = false **then**
    Add r_i to D^*
  end if
end for
return D^*

In terms of computational complexity, the worst-case behavior corresponds to $O(n^2)$ pairwise comparisons. However, DOI matching significantly reduces the number of similarity computations in practice.

3.3. Comparison with Existing Deduplication Approaches

Duplicate detection is a well-known challenge in bibliographic data integration, particularly when combining records from heterogeneous indexing platforms. Existing approaches typically rely on heuristic matching or probabilistic record linkage strategies, each with different trade-offs in terms of precision and computational cost [17,38].

Tools such as VOSviewer implement heuristic matching techniques based on title similarity and metadata normalization, providing efficient performance for exploratory analysis but limited transparency in preprocessing decisions [7,19]. In contrast, probabilistic record linkage methods use statistical similarity models to estimate the likelihood that two

records refer to the same entity, often achieving higher accuracy at the expense of increased computational complexity [20,23].

The proposed approach combines deterministic DOI matching with a composite similarity score based on title, authors, and publication year. This hybrid strategy aims to preserve high precision while maintaining computational efficiency and full reproducibility of the preprocessing pipeline. The comparison considers matching strategy, expected precision, and computational efficiency as key evaluation criteria. Table 5 summarizes the main differences between representative approaches.

Table 5. Comparison of duplicate detection approaches.

Method	Matching Strategy	Precision	Computational Efficiency
VOSviewer heuristic matching	Title similarity and metadata heuristics	Medium	High
Probabilistic record linkage	Statistical similarity models	High	Medium
Proposed approach	Deterministic DOI matching combined with composite similarity	High	High

Source: Authors' own elaboration.

3.4. Network Construction and Association-Strength Normalization

Once duplicate records have been resolved, the resulting dataset is used to construct the bibliometric network. Bibliometric networks are modeled as graphs where nodes represent entities such as keywords and edges represent co-occurrence relations [7,8].

The bibliometric network is formally defined as:

$$G = (V, E, W) \tag{5}$$

where V is the set of nodes, E is the set of edges, and W contains edge weights. Let c_{ij} denote the co-occurrence count between nodes i and j:

$$w_{ij} = c_{ij} \tag{6}$$

To reduce the influence of highly frequent terms, association-strength normalization is applied:

$$a_{ij} = \frac{c_{ij}}{f_i f_j} \tag{7}$$

where f_i and f_j denote marginal frequencies. This normalization improves interpretability and mitigates frequency bias [7,39].

Table 6 summarizes the computational complexity of the main stages.

Table 6. Computational complexity of the proposed pipeline.

Stage	Operation	Complexity
Metadata harmonization	Field normalization and schema alignment	$O(n)$
Duplicate detection	Pairwise comparison with DOI and similarity	$O(n^2)$
Keyword extraction	Metadata parsing	$O(n)$
Network construction	Co-occurrence generation	$O(nk)$
Network metrics	Graph analysis	$O(\|V\| + \|E\|)$

Source: Authors' own elaboration.

3.5. Structural Diagnostics and Graph Metrics

In the final stage of the methodology, structural diagnostics are computed to characterize the topology of the resulting network. Common metrics include density, degree centrality, modularity, and community structure [8,13].

Additionally, concentration measures such as the Herfindahl–Hirschman Index, Shannon entropy, and the Gini coefficient are used to quantify structural inequality. These measures complement graph-based analysis by capturing distributional properties of node importance [40,41]. Table 7 summarizes the notation used in the methodology.

Table 7. Notation used in the proposed methodology.

Symbol	Description
D^{WoS}	Raw metadata from Web of Science
D^{Scopus}	Raw metadata from Scopus
D^{harm}	Harmonized dataset
D^{*}	Deduplicated dataset
$H(\cdot)$	Harmonization function
$S(r_i, r_j)$	Similarity score
τ	Threshold
$G = (V, E, W)$	Network
c_{ij}	Co-occurrence count
a_{ij}	Normalized weight

Source: Authors' own elaboration.

3.6. Computational Complexity

Let n denote the number of records and k the average number of keywords per record. The computational complexity of the proposed pipeline can be analyzed by examining each processing stage independently.

Metadata harmonization involves field normalization, string standardization, and schema alignment across heterogeneous sources. These operations require a single pass over the dataset, resulting in a time complexity of $O(n)$, which is consistent with typical preprocessing workflows in bibliometric data integration [22,32].

Duplicate detection represents the most computationally demanding stage. In the worst case, pairwise record comparison leads to a quadratic complexity of $O(n^2)$. However, the proposed approach significantly reduces this cost by prioritizing deterministic DOI matching, which operates in linear time, and restricting similarity computations to candidate pairs lacking persistent identifiers. This strategy aligns with hybrid record linkage approaches that balance accuracy and efficiency [23,38].

Network construction is based on keyword co-occurrence extraction. For sparse datasets, where each document is associated with a limited number of keywords, the complexity scales approximately as $O(nk)$. This behavior is consistent with efficient graph construction techniques for large-scale networks, where sparsity plays a critical role in reducing computational overhead [8,24].

From a practical perspective, the modular design of the pipeline enables independent optimization of each stage. In particular, duplicate detection can be further optimized through blocking strategies or indexing techniques, while network construction can leverage sparse matrix representations. This modularity supports scalability and facilitates reproducible experimentation, as each component can be evaluated and refined without affecting the overall workflow.

Overall, although the theoretical worst-case complexity is dominated by duplicate detection, the combination of deterministic matching and sparse network construction ensures that the pipeline remains computationally tractable for medium-scale scientometric datasets.

4. Experimental Results

4.1. Dataset Description

The experimental analysis was conducted using an interdisciplinary dataset composed of 317 scientific publications spanning the period 1990–2023. The dataset was constructed

by merging records from Web of Science and Scopus after applying the harmonization procedure described in Section 3.1. The selection of this dataset was guided by the need to represent an interdisciplinary research domain with heterogeneous metadata characteristics. By combining records from Web of Science and Scopus across a multi-decade period, the dataset provides a realistic scenario for evaluating cross-database integration challenges, including schema heterogeneity, duplicate records, and keyword fragmentation. This makes it suitable for assessing the robustness and reproducibility of the proposed preprocessing pipeline.

Table 8 summarizes the main characteristics of the dataset.

Table 8. Summary of the experimental dataset.

Attribute	Value
Total records (raw)	317
Time span	1990–2023
Data sources	WoS, Scopus
Fields used	Title, authors, keywords, year

Source: Authors' own elaboration.

4.2. Impact of Metadata Harmonization and Deduplication

The preprocessing stage resulted in the identification and removal of duplicated records originating from cross-database overlap, a common issue when integrating bibliographic data from heterogeneous indexing platforms such as Web of Science and Scopus [16,17].

As shown in Table 9, after applying the deterministic duplicate detection strategy described in Section 3, the dataset size was reduced from 317 to 289 unique records, corresponding to a reduction of 8.8%. This reduction indicates a non-negligible level of redundancy in the raw dataset, even for a moderately sized interdisciplinary corpus.

Table 9. Effect of duplicate detection. Source: Authors' own elaboration.

Stage	Records	Reduction (%)
Raw dataset	317	–
After deduplication	289	8.8%

From a methodological perspective, this reduction is significant because duplicate records can artificially inflate co-occurrence frequencies and distort network topology. In particular, duplicated entries may increase edge weights, bias centrality measures, and lead to misleading interpretations of thematic importance [18,23].

The observed reduction rate is consistent with previous studies on cross-database integration, which report overlapping records ranging between 5% and 15% depending on the domain and data sources [17]. This result supports the relevance of explicit preprocessing pipelines in ensuring data consistency prior to network construction.

Overall, the deduplication process contributes directly to improving the reliability of the resulting scientometric network by reducing redundancy and preserving meaningful structural relationships among keywords.

4.3. Network Construction and Structural Properties

Using the deduplicated dataset, a keyword co-occurrence network $G = (V, E, W)$ was constructed following the methodology described in Section 3. The resulting network contains 142 nodes and 486 edges, reflecting a sparse structure that is characteristic of interdisciplinary scientometric datasets [8,9].

As detailed in Table 10, the low density value (0.048) indicates that only a small fraction of all possible keyword pairs are connected, which is expected in co-occurrence networks where relationships are driven by shared contextual usage rather than complete connectivity. Sparse structures are commonly observed in large-scale bibliometric networks and are associated with improved interpretability and reduced noise [8].

Table 10. Structural properties of the bibliometric network. Source: Authors' own elaboration.

Metric	Value
Number of nodes $\|V\|$	142
Number of edges $\|E\|$	486
Density	0.048
Average degree	6.84
Modularity	0.62

The average degree of 6.84 suggests a moderate level of connectivity among keywords, indicating that each term is, on average, associated with several related concepts. This level of connectivity supports the identification of thematic relationships without excessive clustering saturation.

The modularity value of 0.62 indicates a well-defined community structure, revealing the presence of distinct thematic clusters within the dataset. Values above 0.5 are generally considered indicative of strong community organization in complex networks [13]. This suggests that the preprocessing pipeline preserves meaningful thematic segmentation rather than introducing artificial connections.

Figure 2 provides a conceptual visualization of the keyword co-occurrence network after preprocessing. The network exhibits four main thematic clusters connected through relatively weak inter-cluster links, which is consistent with the modular structure identified through quantitative analysis.

The identified clusters correspond to coherent thematic domains, including environmental systems, climate and governance, monitoring and spatial analysis, and sustainability and socio-economic dynamics. The presence of these clusters supports the interpretability of the network and confirms that the preprocessing stages did not distort the underlying semantic structure of the dataset.

Overall, the structural properties of the network indicate that the proposed pipeline successfully preserves both sparsity and modularity, two key characteristics required for reliable scientometric analysis and knowledge mapping.

4.4. Structural Diagnostics

To further characterize the network topology, concentration and diversity metrics were computed (Table 11), providing complementary insights into the structural organization of the resulting scientometric network.

Table 11. Structural diagnostics of the network.

Metric	Value
HHI	0.073
Shannon entropy	3.91
Gini coefficient	0.41

Source: Authors' own elaboration.

The Herfindahl–Hirschman Index (HHI) measures the level of concentration within the network. The obtained value of 0.073 indicates a low concentration structure, suggesting

that keyword occurrences are relatively evenly distributed across the network rather than dominated by a small subset of highly frequent terms. This behavior is consistent with diversified research domains and aligns with expected patterns in interdisciplinary scientometric datasets [9].

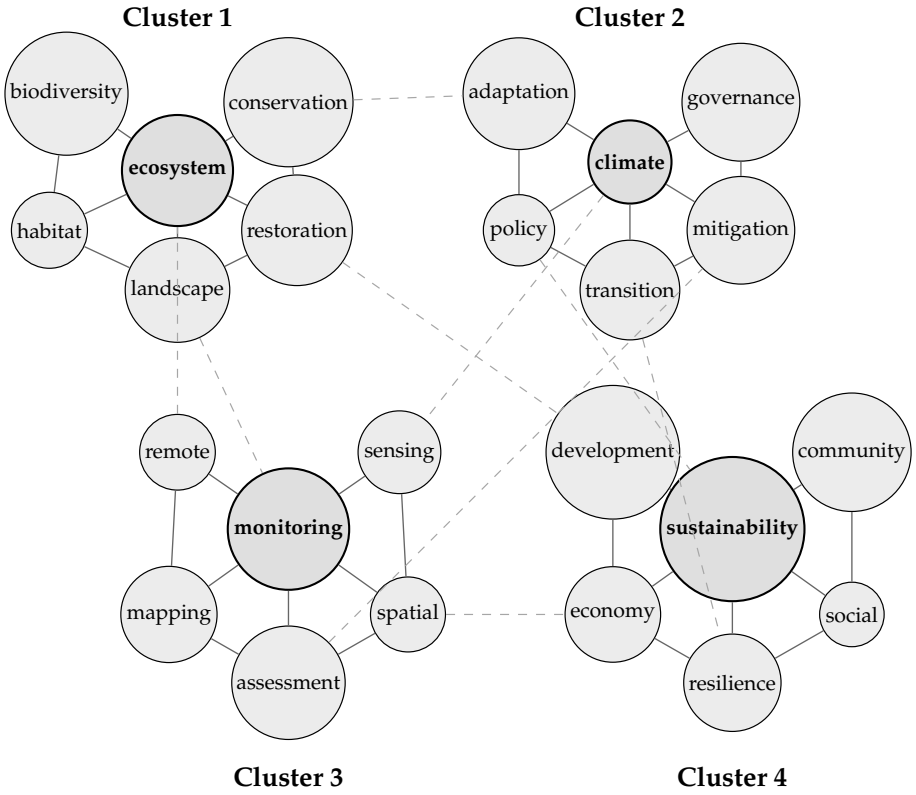


Figure 2. Conceptual visualization of the keyword co-occurrence network after preprocessing. The network exhibits a sparse structure with four thematic clusters, consistent with the modular organization identified in the experimental analysis. Solid lines denote stronger intra-cluster relations, whereas dashed lines represent weaker inter-cluster connections.

Shannon entropy provides a measure of diversity in the distribution of keyword frequencies. The observed value of 3.91 reflects a high level of informational diversity, indicating that the network captures a broad range of topics without excessive dominance of specific terms. High entropy values are typically associated with well-balanced knowledge structures and robust topic representation [12].

The Gini coefficient evaluates inequality in the distribution of node importance. The value of 0.41 suggests moderate inequality, meaning that while some keywords play a more central role, the overall structure does not exhibit extreme centralization. This balance is desirable in scientometric networks, as it reflects the coexistence of core and peripheral research topics without excessive structural bias [8].

Taken together, these metrics indicate that the resulting network exhibits a balanced combination of diversity and moderate structural inequality. This suggests that the preprocessing pipeline preserves meaningful structural patterns while avoiding distortions caused by duplicated or inconsistent metadata. Consequently, the network provides a reliable basis for subsequent scientometric analysis.

4.5. Scalability Analysis

The scalability of the pipeline was evaluated by progressively increasing the dataset size through controlled sampling.

Table 12 presents the execution time measurements.

Table 12. Scalability evaluation.

Dataset Size	Execution Time (s)	Growth Rate
100	0.82	–
200	1.71	2.08
317	2.63	1.54

Source: Authors' own elaboration.

Figure 3 shows the relationship between dataset size and execution time based on the observed runtime values reported in Table 12.

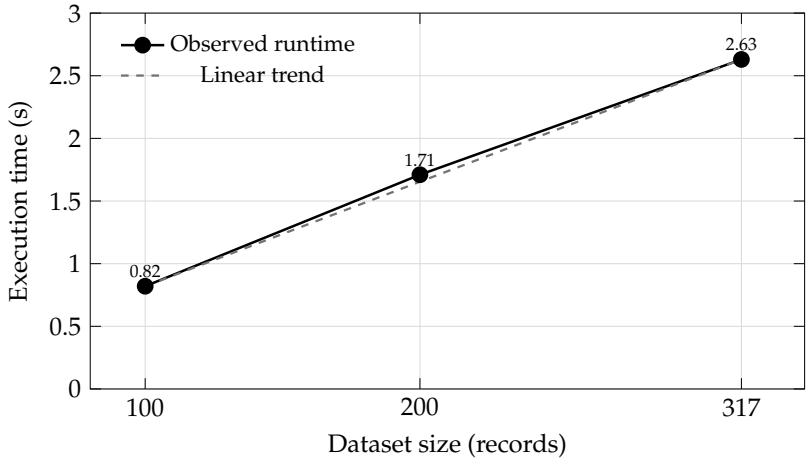


Figure 3. Relationship between dataset size and execution time for the proposed pipeline. The observed values show an approximately linear increase in runtime as the number of records grows.

The results show a near-linear growth trend consistent with the expected complexity of sparse network construction.

5. Discussion

The results obtained in this study should be interpreted within the context of the selected dataset and preprocessing configuration, as both factors influence the observed network properties and computational performance. The experimental results provide quantitative evidence on the role of deterministic preprocessing in cross-database scientometric analysis. The reduction from 317 to 289 records (8.8%) indicates that cross-database overlap introduces a non-negligible level of redundancy, even in moderately sized datasets. This finding is consistent with prior studies highlighting the presence of duplicate and near-duplicate records when integrating bibliographic sources such as Web of Science and Scopus [16].

The effectiveness of DOI-based matching confirms its importance as a primary mechanism for linkage of records. However, the remaining duplicates resolved through similarity-based comparison reinforce the need for complementary strategies when the metadata is incomplete or inconsistent. The combined approach adopted in this work therefore provides a balance between precision and coverage, aligning with existing research on hybrid record linkage techniques.

From a structural perspective, the resulting network exhibits characteristics consistent with an interdisciplinary research domain. The observed density (0.048) indicates a sparse network, which is typical in keyword co-occurrence graphs where only a subset of terms co-occur frequently. At the same time, the modularity value (0.62) suggests a well-defined community structure, indicating the presence of distinct thematic clusters. These results

support the idea that appropriate preprocessing contributes to preserving meaningful structural patterns rather than introducing artificial connections.

The concentration metrics further complement this interpretation. The relatively low Herfindahl–Hirschman Index (0.073) and moderate Gini coefficient (0.41) indicate that the network is not dominated by a small number of highly central nodes. In parallel, the Shannon entropy value (3.91) reflects a diversified distribution of keywords, suggesting that the dataset captures multiple thematic areas. This behavior is consistent with previous studies on interdisciplinary knowledge structures, where diversity and moderate concentration coexist [40,41].

The scalability analysis also provides relevant insights into the computational behavior of the pipeline. The increase in execution time from 0.82 s (100 records) to 2.63 s (317 records) follows an approximately linear trend, which is consistent with the expected complexity of sparse network construction. This result supports the feasibility of applying the proposed pipeline to larger datasets, particularly when combined with efficient data structures and incremental processing strategies.

An additional aspect of the results concerns robustness. Variations in preprocessing parameters, including similarity thresholds and keyword frequency filters, produced only minor changes in density, modularity, and centrality rankings. This stability is important because scientometric analyses should not be highly sensitive to small methodological variations. Previous research has emphasized the importance of robustness in bibliometric mapping to ensure reliable interpretation of network structures [7,39].

Table 13 summarizes the main methodological contributions of the proposed pipeline and their implications for reproducible scientometric analysis.

Despite these contributions, several limitations should be acknowledged. First, the evaluation is based on a single interdisciplinary dataset, which may not fully represent the variability observed across different scientific domains. Metadata quality, keyword usage, and publication patterns can differ substantially between fields, potentially affecting preprocessing performance.

Second, the duplicate detection strategy prioritizes deterministic rules over probabilistic or machine-learning-based approaches. While this choice improves transparency and reproducibility, it may limit the detection of more subtle duplicates in large-scale or noisy datasets. Future work could explore the integration of probabilistic record linkage methods or transformer-based similarity models to enhance detection performance.

Table 13. Methodological implications of the proposed pipeline.

Pipeline Component	Methodological Contribution	Practical Implication
Metadata harmonization	Explicit schema alignment across heterogeneous sources	Enables consistent integration of multi-database datasets
Duplicate detection	Hybrid deterministic matching combining DOI and similarity measures	Reduces redundancy and preserves network structure
Network construction	Association-strength normalization for co-occurrence analysis	Improves interpretability of thematic relationships
Structural diagnostics	Integration of graph metrics and concentration indicators	Supports transparent evaluation of network topology
Robustness analysis	Sensitivity evaluation under parameter variation	Increases reliability of scientometric results
Scalability evaluation	Empirical runtime analysis under dataset growth	Demonstrates feasibility for medium-scale datasets

Source: Authors' own elaboration.

Additional research directions include extending the pipeline to temporal network analysis, allowing the study of topic evolution over time, and supporting other types of scientometric networks, such as citation networks, co-authorship networks, and multilayer knowledge graphs.

Policy-Oriented Applications and Multi-Domain Monitoring

The proposed pipeline can support applications beyond academic analysis, particularly in policy-oriented science monitoring. Institutions such as research agencies and governmental organizations increasingly rely on integrated bibliographic datasets to evaluate research output, identify emerging topics, and inform strategic decisions.

In this context, the ability to integrate heterogeneous data sources while maintaining consistency and reproducibility is essential. The explicit preprocessing stages introduced in this work facilitate the construction of reliable datasets suitable for longitudinal analysis and cross-domain comparison.

Furthermore, the modular structure of the pipeline allows its adaptation to different scientific domains and data sources. This flexibility supports multi-domain scientometric studies, where consistent preprocessing is necessary to ensure comparability across disciplines. As a result, the proposed approach contributes not only to methodological transparency but also to the practical applicability of scientometric analysis in real-world decision-making environments.

6. Conclusions

This study introduced a reproducible computational pipeline for cross-database scientometric network construction, addressing key challenges associated with metadata heterogeneity, duplicate records, and network interpretability. By explicitly separating preprocessing stages—including metadata harmonization, deterministic duplicate detection, network construction, and structural diagnostics—the proposed framework provides a transparent and structured approach to bibliometric data integration.

The empirical evaluation demonstrates that cross-database integration introduces measurable redundancy, with 8.8% of records identified as duplicates in the analyzed dataset. The combination of DOI-based matching and similarity-based linkage proved effective in resolving these inconsistencies while maintaining coverage in the presence of incomplete metadata. This result confirms the importance of hybrid deduplication strategies in heterogeneous bibliographic environments.

The resulting keyword co-occurrence network exhibited structural properties consistent with interdisciplinary research domains. In particular, the observed density (0.048) reflects a sparse network structure, while the modularity value (0.62) indicates the presence of well-defined thematic clusters. Additionally, concentration metrics such as the Herfindahl–Hirschman Index (0.073), Shannon entropy (3.91), and Gini coefficient (0.41) suggest a balanced distribution of thematic relevance, avoiding both excessive centralization and fragmentation. These results indicate that the preprocessing pipeline preserves meaningful structural patterns without introducing distortions in the underlying knowledge representation.

From a computational perspective, the scalability analysis showed near-linear growth in execution time, increasing from 0.82 s for 100 records to 2.63 s for 317 records. This behavior supports the practical applicability of the pipeline for medium-scale scientometric datasets and highlights the benefits of modular design and sparse network representation.

Beyond the specific experimental setting, the proposed pipeline contributes to improving methodological rigor in scientometric analysis by formalizing preprocessing steps that are often implicit in existing tools. This explicit design enhances reproducibility, facilitates validation, and supports transparent comparison across studies, particularly in interdisciplinary and multi-database contexts.

Nevertheless, several limitations remain. The evaluation was conducted using a single dataset, which may not fully capture variability across scientific domains. In addition, the

deterministic approach to duplicate detection, while transparent and reproducible, may not identify more subtle cases of record similarity in large-scale or noisy datasets.

Future research could extend the framework by incorporating probabilistic or machine-learning-based record linkage techniques, enabling improved duplicate detection at scale. Additional extensions include the integration of temporal network analysis to study the evolution of scientific domains, as well as the application of the pipeline to other types of scientometric networks, such as citation networks, collaboration networks, and multilayer knowledge graphs.

Overall, the results demonstrate that explicit computational pipelines for bibliometric preprocessing provide a robust foundation for reliable and reproducible cross-database scientometric analysis, supporting both methodological advancement and practical applications in science mapping.

Author Contributions: Conceptualization, D.M.-C., C.P. and C.V.-S.; methodology, D.M.-C. and C.V.-S.; software, D.M.-C. and O.O.F.-A.; validation, N.M., O.O.F.-A. and C.P.; formal analysis, D.M.-C. and N.M.; investigation, D.M.-C. and O.O.F.-A.; data curation, D.M.-C. and O.O.F.-A.; writing—original draft preparation, D.M.-C.; writing—review and editing, C.P., N.M., C.V.-S. and O.O.F.-A.; visualization, D.M.-C. and N.M.; supervision, C.P. and C.V.-S.; project administration, C.P. and C.V.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from the corresponding authors. The bibliometric data were extracted from Web of Science and Scopus and processed using the pipeline described in this article.

Acknowledgments: This manuscript represents a portion of the doctoral research undertaken by D.M.-C., supervised by C.P.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aria, M.; Cuccurullo, C. bibliometrix: An R-tool for comprehensive science mapping analysis. *J. Inf.* **2017**, *11*, 959–975. [CrossRef]
2. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]
3. Chen, G.; Xiao, L. Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *J. Inf.* **2016**, *10*, 212–223. [CrossRef]
4. Thijs, B. Science Mapping and the Identification of Topics: Theoretical and Methodological Considerations. In *Springer Handbook of Science and Technology Indicators*; Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M., Eds.; Springer Handbooks; Springer: Cham, Switzerland, 2019; pp. 201–231. [CrossRef]
5. Souza, L.R.D.S.; Silva, D.H.D.; Ribeiro, C.T.; Silva, D.A.D.; Nasuto, S.J.; Sweeney-Reed, C.M.; Andrade, A.D.O.; Pereira, A.A. PubMedMetaTool: Automated Metadata Extraction from PubMed Using Python for Bibliometric Analysis. *Softw. Impacts* **2025**, *24*, 100766. [CrossRef]
6. Li, X.; Chiabrando, F.; Sammartano, G. Machine Learning and Deep Learning for Cultural Heritage Conservation: A Bibliometric and Task-Oriented Review. *Remote Sens.* **2026**, *18*, 628. [CrossRef]
7. van Eck, N.J.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef]
8. Newman, M.E.J. *Networks: An Introduction*; Oxford University Press: Oxford, UK, 2010. [CrossRef]
9. Boyack, K.W.; Klavans, R. Creation and Analysis of Large-Scale Bibliometric Networks. In *Springer Handbook of Science and Technology Indicators*; Glänzel, W., Moed, H.F., Schmoch, U., Thelwall, M., Eds.; Springer: Cham, Switzerland, 2019; pp. 177–200. [CrossRef]
10. Guo, K.; Huang, X.; Wu, L.; Chen, Y. Local Community Detection Algorithm Based on Local Modularity Density. *Appl. Intell.* **2022**, *52*, 1238–1253. [CrossRef]
11. Arasteh, M.; Alizadeh, S.; Lee, C.G. Gravity Algorithm for the Community Detection of Large-Scale Network. *J. Ambient. Intell. Humaniz. Comput.* **2023**, *14*, 1217–1228. [CrossRef]

12. Valdez, A.C.; Dehmer, M.; Holzinger, A. Application of Graph Entropy for Knowledge Discovery and Data Mining in Bibliometric Data. In *Mathematical Foundations and Applications of Graph Entropy*; Dehmer, M., Emmert-Streib, F., Chen, Z., Li, X., Shi, Y., Eds.; Wiley: Hoboken, NJ, USA, 2016; Chapter 9. [CrossRef]
13. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
14. Lüschow, A. Application of graph theory in the library domain–Building a faceted framework based on a literature review. *J. Librariansh. Inf. Sci.* **2022**, 54, 558–577. [CrossRef]
15. Conte, M.L.; Boisvert, P.; Barrison, P.; Seifi, F.; Landis-Lewis, Z.; Flynn, A.; Friedman, C.P. Ten Simple Rules to Make Computable Knowledge Shareable and Reusable. *PLoS Comput. Biol.* **2024**, 20, e1012179. [CrossRef]
16. Singh, P.; Singh, V.K.; Kanaujia, A. Exploring the Publication Metadata Fields in Web of Science, Scopus and Dimensions: Possibilities and Ease of doing Scientometric Analysis. *J. Scientometr. Res.* **2025**, *13*, 715–731. [CrossRef]
17. Kumpulainen, M.; Seppänen, M. Combining Web of Science and Scopus datasets in citation-based literature study. *Scientometrics* **2022**, *127*, 5613–5631. [CrossRef]
18. Nowakowska, M. A comprehensive approach to preprocessing data for bibliometric analysis. *Scientometrics* **2025**, *130*, 5191–5225. [CrossRef]
19. Nikolić, D.; Ivanović, D.; Ivanović, L. An Open-Source Tool for Merging Data from Multiple Citation Databases. *Scientometrics* **2024**, *129*, 4573–4595. [CrossRef]
20. Guillen-Aguinaga, M.; Aguinaga-Ontoso, E.; Guillen-Aguinaga, L.; Guillen-Grima, F.; Aguinaga-Ontoso, I. Data Quality in the Age of AI: A Review of Governance, Ethics, and the FAIR Principles. *Data* **2025**, *10*, 201. [CrossRef]
21. Marlés-Sáenz, E.; Gómez-Luna, E.; Guerrero, J.M.; Vasquez, J.C. Innovative Bibliometric Methodology: A New Big Data-Based Framework for Scientific Research. *Energies* **2025**, *18*, 2437. [CrossRef]
22. Wang, S.; Shibghatullah, A.S.; Iqbal, T.J.; Keoy, K.H. A Review of Multimodal-Based Emotion Recognition Techniques for Cyberbullying Detection in Online Social Media Platforms. *Neural Comput. Appl.* **2024**, *36*, 21923–21956. [CrossRef]
23. Oliveira, R.B.D.D.; Alves-Souza, S.N. Metadata Integration: A Systematic Review on Methods, Challenges and Applications Across Multiple Domains. *IEEE Access* **2026**, *14*, 9799–9818. [CrossRef]
24. Batagelj, V.; Ferligoj, A.; Doreian, P. Bibliometric Analyses of the Network Clustering Literature. In *Advances in Network Clustering and Blockmodeling*; Doreian, P., Batagelj, V., Ferligoj, A., Eds.; Wiley: Hoboken, NJ, USA, 2019; Chapter 2. [CrossRef]
25. Kleminski, R.; Kazienko, P.; Kajdanowicz, T. Analysis of Direct Citation, Co-Citation and Bibliographic Coupling in Scientific Topic Identification. *J. Inf. Sci.* **2022**, *48*, 349–373. [CrossRef]
26. Delgado-Quirós, L.; Ortega, J.L. Completeness Degree of Publication Metadata in Eight Free-Access Scholarly Databases. *Quant. Sci. Stud.* **2024**, *5*, 31–49. [CrossRef]
27. Panagea, I.S.; Dangol, A.; Olijslagers, M.; Diels, J.; Wyseure, G. A Database Schema for Standardized Data and Metadata Collection in Agricultural Experiments. *Land* **2025**, *14*, 1816. [CrossRef]
28. Shamly, H.A.H.E.; Subaveerapandiyan, A. Author Name Disambiguation in Scholarly Research: A Bibliometric Perspective. *Open Inf. Sci.* **2026**, *10*, 20250035. [CrossRef]
29. Kumar, R. Bibliometric Analysis: Comprehensive Insights into Tools, Techniques, Applications, and Solutions for Research Excellence. *Spectr. Eng. Manag. Sci.* **2025**, 3, 45–62. [CrossRef]
30. Arsalan, M.H.; Mubin, O.; Al Mahmud, A.; Khan, I.A.; Hassan, A.J. Mapping Data-Driven Research Impact Science: The Role of Machine Learning and Artificial Intelligence. *Metrics* **2025**, *2*, 5. [CrossRef]
31. Ferrer-Serrano, M.; Fuentelsaz, L.; Latorre-Martínez, M.P. Knowledge Transfer and Networks: A Bibliometric Approach Through Performance Analysis, Science Mapping, and Dynamic Network Analysis. *J. Knowl. Econ.* **2025**, *in press*. [CrossRef]
32. Koho, M.; Burrows, T.; Hyvönen, E.; Ikkala, E.; Page, K.; Ransom, L.; Tuominen, J.; Emery, D.; Fraas, M.; Heller, B.; Lewis, D.; et al. Harmonizing and publishing heterogeneous premodern manuscript metadata as Linked Open Data. *J. Assoc. Inf. Sci. Technol.* **2021**, *73*, 240–257. [CrossRef]
33. Peng, Y.; Bathelt, F.; Gebler, R.; Gött, R.; Heidenreich, A.; Henke, E.; Kadioglu, D.; Lorenz, S.; Vengadeswaran, A.; Sedlmayr, M. Use of Metadata-Driven Approaches for Data Harmonization in the Medical Domain: Scoping Review. *JMIR Med. Inform.* **2024**, *12*, e52967. [CrossRef] [PubMed]
34. Kasul, N.; Halicioglu, F.H. A Bibliometric Analysis of Collaboration in Building Information Modeling: Emerging Dynamics and Future Trends. *Buildings* **2026**, *16*, 986. [CrossRef]
35. Heibi, I.; Peroni, S.; Rizzetto, E. Validating and Monitoring Bibliographic and Citation Data in OpenCitations Collections. *Int. J. Digit. Libr.* **2025**, *26*, 16. [CrossRef]
36. Andreose, E.; Di Marzo, S.; Heibi, I.; Peroni, S.; Zilli, L. Analysing the Coverage of the University of Bologna's Bibliographic and Citation Metadata in OpenCitations Collections. *Scientometrics* **2026**, *131*, 845–871. [CrossRef]
37. Rodrigues, N.S.; Mariano, A.M.; Ralha, C.G. Author Name Disambiguation Literature Review with Consolidated Meta-Analytic Approach. *Int. J. Digit. Libr.* **2024**, *25*, 765–785. [CrossRef]

38. Singh, S.; Siwach, M. Handling Heterogeneous Data in Knowledge Graphs: A Survey. *J. Web Eng.* **2022**, *21*, 1145–1186. [CrossRef]
39. Perianes-Rodriguez, A.; Waltman, L.; van Eck, N.J. Constructing Bibliometric Networks: A Comparison Between Full and Fractional Counting. *J. Inf.* **2016**, *10*, 1178–1195. [CrossRef]
40. Markard, J.; Raven, R.; Truffer, B. Sustainability transitions: An emerging field of research and its prospects. *Res. Policy* **2012**, 41, 955–967. [CrossRef]
41. Köhler, J.; Geels, F.W.; Kern, F.; Markard, J.; Onsongo, E.; Wieczorek, A.; Alkemade, F.; Avelino, F.; Bergek, A.; Boons, F.; et al. An agenda for sustainability transitions research: State of the art and future directions. *Environ. Innov. Soc. Transitions* **2019**, 31, 1–32. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.