Network Analysis of Research Citations

Kiana Seraj

Novmber 2024

1 Introduction

1.1 Introduction to Networks

A network is a powerful representation that is used to model relationships between entities across various disciplines. It consists of nodes (representing entities) and edges (indicating relationships between nodes). Networks are incredibly versatile and are used to study complex systems that range from social interactions and computer networks to biological systems and transportation routes. Networks allow us to uncover connectivity patterns that are not apparent in isolated data points. By analyzing how entities interact, we can understand their roles, identify key influencers, and detect clusters of closely related entities. Such analyses are crucial in numerous fields, where connections among individuals, institutions, or ideas define the behavior of the entire system[1].

1.2 Citation Networks and Their Properties

A citation network is a specialized type of network used to map academic literature and how knowledge is propagated within a scientific field. In a citation network, the nodes represent research papers, and the edges are directed links showing when one paper cites another. Citation networks provide a fascinating glimpse into the evolution of research and the flow of knowledge between academic works. Key properties of citation networks include:

Directed Nature: Citation networks are inherently directed because each edge has a direction—from the citing paper to the cited paper. This structure highlights the lineage of ideas and knowledge across years of research.

Centrality Measures: Metrics such as degree centrality and betweenness centrality help determine the influence of specific papers. Highly central papers are often pivotal in shaping the direction of research, being either foundational studies or critical reviews that are frequently cited. Analyzing such properties provides information on the impact of individual papers and how ideas spread across different domains[2].

1.3 Aim of the Project

The primary aim of this project is to analyze a sampled citation network to identify central and influential papers within the field. By leveraging network analysis techniques, the study seeks to uncover key papers that act as knowledge hubs, are highly cited, and play a pivotal role in shaping the research landscape. The analysis also focuses on detecting cliques, understanding the distribution of in-degree and out-degree nodes, and examining specific characteristics of the papers, such as commonly used keywords and prominent journals. Furthermore, the project aims to discern patterns of on-topic papers related to antimicrobial resistance (AMR) and drug resistance within the sampled citation network, shedding light on the multidisciplinary nature of the field and its research trends.

2 Methodology

2.1 Data

The dataset used in this study represents a citation network, where nodes correspond to research papers, and edges represent citation relationships between them. The primary dataset includes an edge list with 479,802 citations and 1,019,630 edges, detailing the citation connections. A supplementary dataset provides metadata for 235,871 of these nodes, offering additional information about the papers, such as journal titles, publication years, and keywords.
Given the size and complexity of the network, computational constraints necessitated sampling a fraction of nodes for certain analyses. To ensure data quality, preprocessing tasks were performed, including handling missing values and removing duplicates. These steps were critical to maintaining the integrity of the dataset and enabling effective network analysis.

2.2 node features

To better understand the characteristics of the papers within the citation network, an analysis of journal titles and keywords frequently used from 1965 to 2023 was conducted. The five most frequently occurring journal titles and keywords were identified. This analysis provides insight into the dominant themes, research trends, and publication outlets within the dataset. By examining these features, the study aims to capture the focus areas and thematic evolution of research over time.

2.3 Community Detection with Louvain method

To identify communities within the large directed network, the Louvain community detection method was applied. This method, a greedy optimization algorithm, is particularly well-suited for detecting non-overlapping communities in large-scale networks [3].
As the Louvain method is designed for undirected graphs, the directed network was first converted to an undirected format. The algorithm then identified 276 distinct communities. These communities reveal the underlying structure of the network, grouping nodes based on their connectivity patterns.
Additionally, in-degree and out-degree values are computed:
In-degree: The number of edges directed toward a node, highlighting nodes that are frequently referenced.
Out-degree: The number of edges directed away from a node, indicating nodes that actively reference others.

2.4 Network metrics

1.Degree Centrality: It measures the proportion of nodes a given node is connected to. In an undirected graph, it simply counts the number of connections. In a directed graph, you can have in-degree (number of incoming edges) and out-degree (number of outgoing edges) centrality. Degree Centrality ranges from 0 to 1.
0: Node has no connections.
1: Node is connected to all other nodes in the network.
The formula for degree centrality is:

$$C_D(v) = \frac{\deg(v)}{n-1}$$

Where:

- C_D: Degree centrality of node v.
- $\deg(v)$: Number of edges connected to node v.
- n: Total number of nodes in the graph.

2.Betweenness Centrality: It quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It reflects the node's role in facilitating communication within the network. Betweenness Centrality ranges from 0 to a theoretical maximum based on the network's structure.
0: Node does not lie on any shortest path.
Higher Values: Nodes that frequently lie on shortest paths between other nodes.

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where:

- $C_B(v)$: Betweenness centrality of node v.
- σ_{st}: Total number of shortest paths from node s to node t.
- $\sigma_{st}(v)$: Number of those shortest paths that pass through node v.

3.Clique: A clique in a network is a subset of nodes where every two distinct nodes are connected by a unique edge. Essentially, it's a completely connected subgraph. Characteristics:
Maximal Cliques: Cliques that cannot be extended by including an adjacent node.
Clique Size: Number of nodes within the clique.
Number of Cliques: Varies based on network structure; larger networks typically have more cliques.
Clique Size: Ranges from 3 (as per your filtering) up to the total number of nodes in the largest fully connected subset.

$$\text{A clique is a subset of nodes } S \subseteq V \text{ such that } \forall u, v \in S,\ (u, v) \in E.$$

where:

- S: A subset of nodes in the graph.
- V: The set of all nodes in the graph.
- E: The set of all edges in the graph.
- $(u, v) \in E$: Indicates that there is an edge between nodes u and v.

3 result

3.1 Node features

To gain initial insights into the characteristics of the papers, key variables were analyzed and evaluated. This included identifying mostly used journal titles and keywords and the number of papers in the area of bacterial and drug resistancy related theme in the interval of 1965 to 2023. This provides an overview of the dominant themes and trends within the dataset.

- **Journal titles**: The top five most frequently used journal titles for papers published between 1975 and 2023 were identified, offering insights into the primary publication outlets within the field during this period.

Journal title	Counts
'Antimicrobial agents and chemotherapy'	10279
'Journal of bacteriology'	10177
'The Journal of antimicrobial chemotherapy '	5548
'Proceedings of the National Academy of Sciences of the United States of America'	5460
'Journal of clinical microbiology '	4318

- **Top 5 highly used keywords of papers descriptors**: The analysis revealed that prominent research in the dataset primarily focused on data obtained from studies involving humans, followed by animals, and then females, in that order. The central topics of these studies predominantly revolved around antibacterial agents and drug resistance, highlighting the critical areas of focus within the field of antimicrobial resistance research.

Descriptors	Counts
'Humans'	113695
'Drug Resistance, Microbial'	61537
'Anti-Bacterial Agents'	54418
'Animals'	44815
'Female'	35721

- **trend of the topic over years**: The trend analysis of the topic over the years reveals consistent growth starting from 1965, peaking over time, and then showing a decline beginning around 1999. This pattern reflects the evolving interest and research focus within the field, Fig 1.

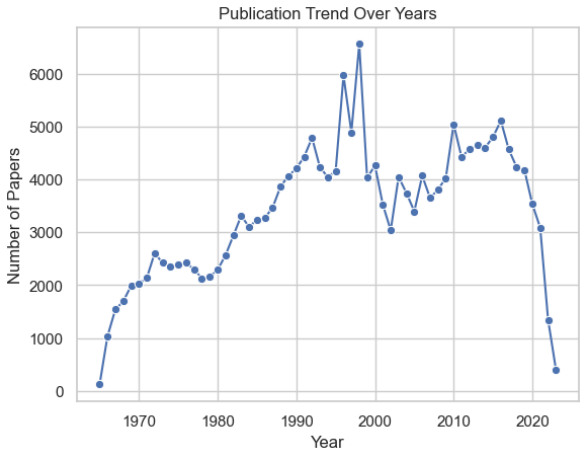

Figure 1: The trend of publishing papers in the field of antimicrobial resistance and drug discovery

3.2 Network characteristics

- **In-degree vs out-degree:**
  Some mostly viewed features within the disributions of in-degree and out-degrees are as follows:
  1.Skewed Distribution (Long-Tail): The in-degree and out-degree distributions exhibit a long-tail or power-law pattern, typical of scale-free networks like citation graphs. Most nodes have low degrees, while a few have very high degrees, Fig 2.
  2.High In-Degree and Out-Degree Nodes: High In-Degree nodes Represent influential or foundational papers cited frequently. High Out-Degree Nodes show Likely review or survey articles referencing many other works.
  3.Scatterplot Insights:
  Concentration Near Origin: Most nodes have low in-degrees and out-degrees, reflecting papers that are rarely cited and cite few others.
  High Out-Degree and Low In-Degree: Review papers citing many but rarely cited themselves.
  High In-Degree and Low Out-Degree: Foundational works widely cited but referencing few others.
  This distribution and scatterplot visualization confirm common trends in real-world networks like citation graphs, revealing the existence of some nodes, papers, being highly popular and the majority being peripheral.

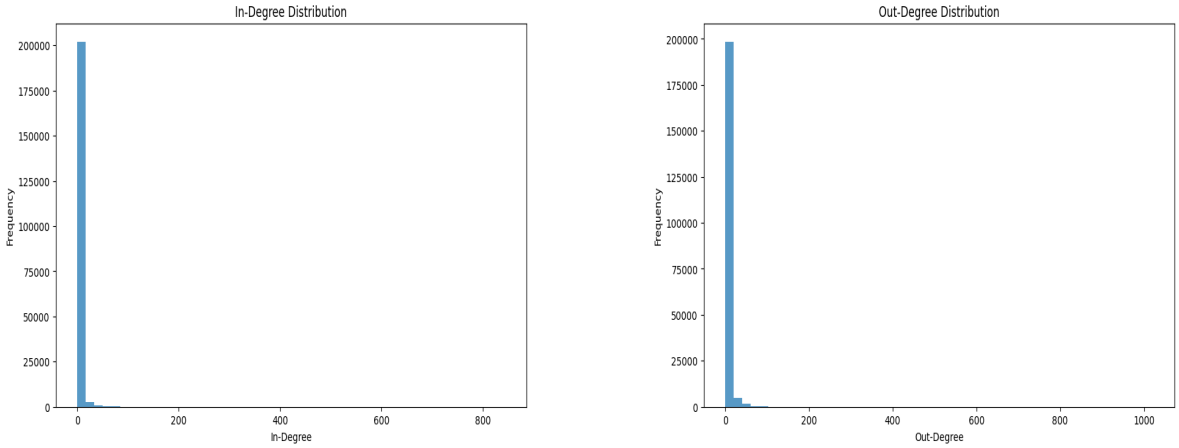

Figure 2: In-Degree and Out-degree counts

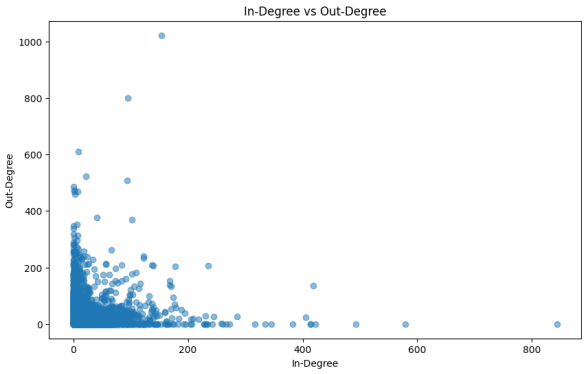

Figure 3: Scatterplot of in-degree vs out-degree

- **Metrics**:
  Some network metrics, such as clique detection, centrality measures, and betweenness centrality, were evaluated to gain insights into the structure and dynamics of the network. These metrics help identify key nodes, measure their influence within the network, and detect closely connected subgroups or cliques, providing a comprehensive understanding of the network's characteristics.
  1.Centrality: Using 479802 nodes with 1019630 edges; the obtained values for centrality is as follows:
  Degree Centrality - Min: 0.0000, Max: 0.0071, Average: 0.0000
  PMID of top 5 nodes having highest centralities are as follows:

PMID	Centrality
'10334980'	0.0071
'5325707'	0.0065
'26476454'	0.0060
'9634230'	0.0060
'6295879'	0.005

  For smaller node size of 200 for the year of 2020, the centrality the Degree Centrality is with:
  Min: 0.0030, Max: 0.0178, Average: 0.0035
  The network representation, scaled by centrality values, illustrates the citation patterns of nodes. It reveals that a node's centrality increases proportionally with the number of citations it makes, highlighting the influence of highly citing papers within the network, Fig 4.

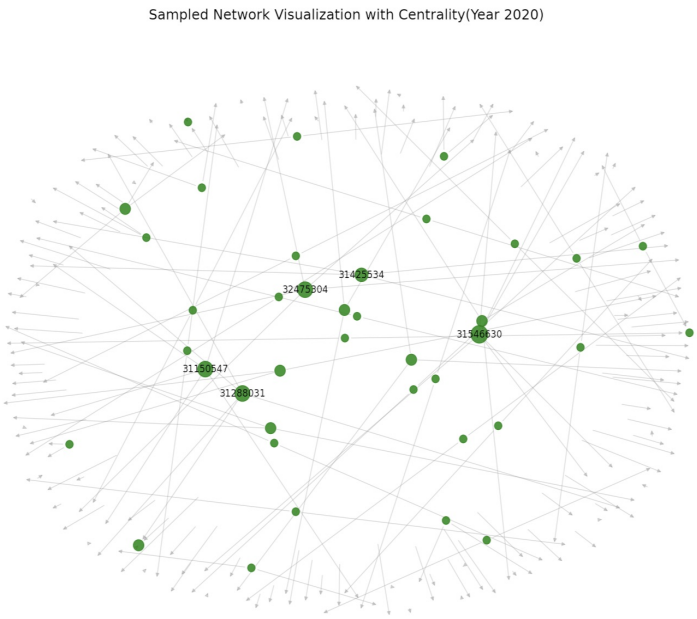

Figure 4: Network representation of 200 nodes, scaled with the centrality

2.Betweenness: Calculating betweenness centrality for all nodes in the network was computationally prohibitive due to the network's large size. To address this, two smaller samples of 200 nodes each, from the years 2010 and 2020, were analyzed. However, both samples exhibited a betweenness centrality of 0 for all nodes, indicating that none of the sampled nodes served as bridges in the shortest paths between other nodes. This suggests a lack of significant intermediary roles for these nodes within their respective samples, possibly due to their isolated positions or the localized nature of the sub-networks analyzed.

3.Cliques: The analysis identified cliques of various sizes in the network, ranging from 2 to 11 nodes, with the largest clique consisting of 11 nodes. An example of the papers included in this largest clique, based on their PubMed IDs (PMIDs), are as follows: [6307967, 7608098, 9333027,

8282690, 8383113, 9068629, 7928997, 2848006, 7504664, 1715857, 8491710]. This demonstrates the presence of highly interconnected subgroups within the network, where every node in a clique is directly connected to every other node, highlighting tightly linked research clusters.

- **community detection:**
  Using the Louvain method, 276 communities were identified within the network. A bar plot (Fig. 5) illustrate the size distribution of these communities, which follows a right-skewed pattern. The largest community contains approximately 24,000 nodes, while the majority of communities are much smaller, with sizes ranging from the hundreds to the tens. This highlights the uneven distribution of community sizes, with a few large, dominant communities and many smaller ones, reflecting the diverse structure of the network.

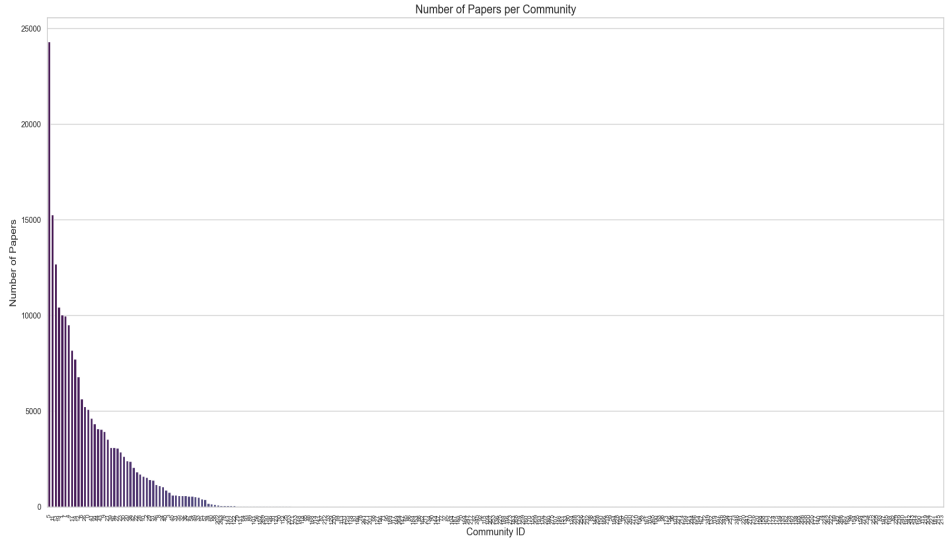

Figure 5: Size of 276 deceted communities

- **proportion of On-topic Papers**:
  Figure 6 illustrates the proportion of on-topic papers within the 276 detected communities. On average, the majority of communities have less than 20% on-topic citations, highlighting the multidisciplinary nature of the research. This suggests that many papers draw significant influence from topics beyond their primary focus, reflecting the interconnected and cross-disciplinary trends in the field.

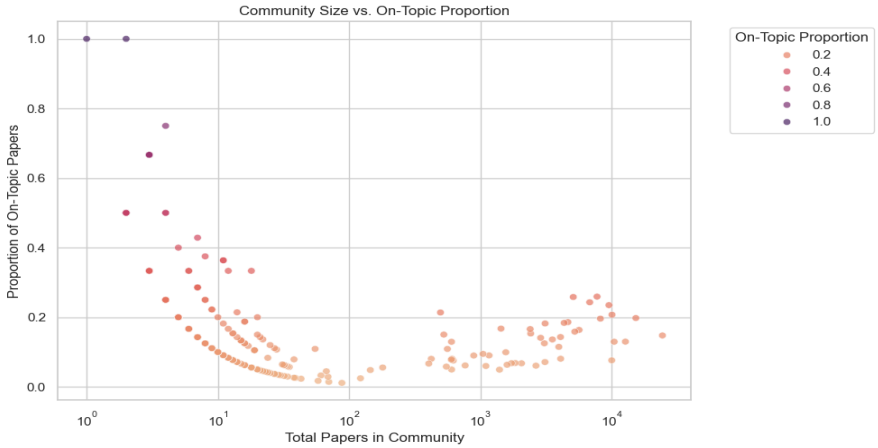

Figure 6: The proportion of on-topic papers in 276 communities based on their sizes

- **Network visualization:**
  To manage the large size of the network, a sample of 200 nodes from the year 2020 was selected for analysis (Fig. 7). The network visualization highlights papers at the center that cite a diverse range of other works, reflecting the broad citation patterns within the network. However, due to the small fraction of nodes chosen, this visualization represents only a limited subset of the overall network, providing a glimpse into its broader structure and dynamics.

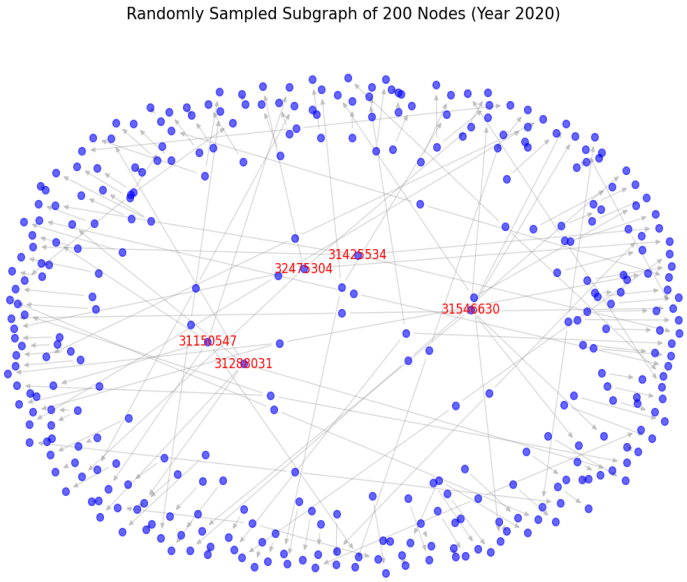

Figure 7: network visualization of 200 nodes sampled of year 2020

- **Network visualization colored by on-topic vs off-topic papers:**
  The network illustrates the widespread citation patterns of papers, including off-topic citations that are indirectly related to the primary focus of the work (Fig. 8). This highlights the multidisciplinary nature of the field and underscores the significant influence of scientific results from diverse areas on the research, demonstrating how interconnected and interdependent various domains are within the citation network.

- **Journal title counts in the sampled 200 nodes network:**
  The bar plot (Fig. 9) reveals that within the network structure of the 200-node sample, the most frequently cited journals include PLOS ONE, Scientific Reports, and the International Journal of Environmental Research and Public Health. These journals stand out as the most commonly referenced publication outlets, reflecting their prominence and influence within the sampled network.

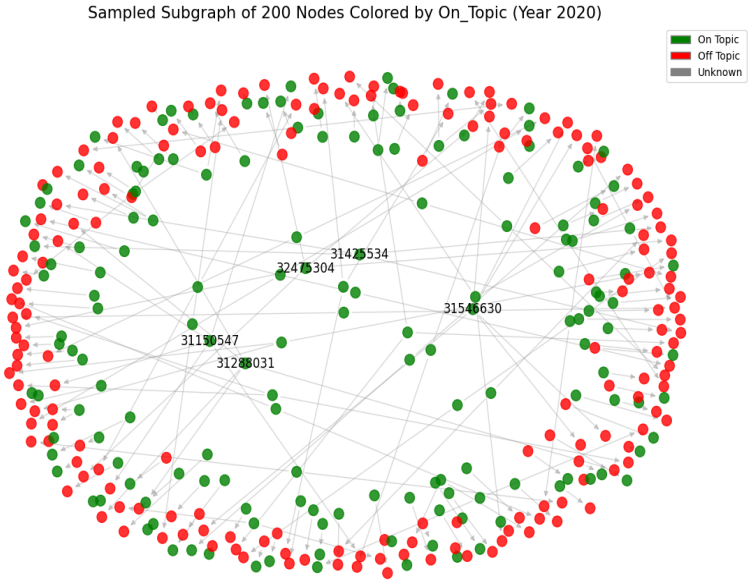

Figure 8: network visualization of 200 nodes sampled of year 2020, colored by boolean values of on-topic, off-topic

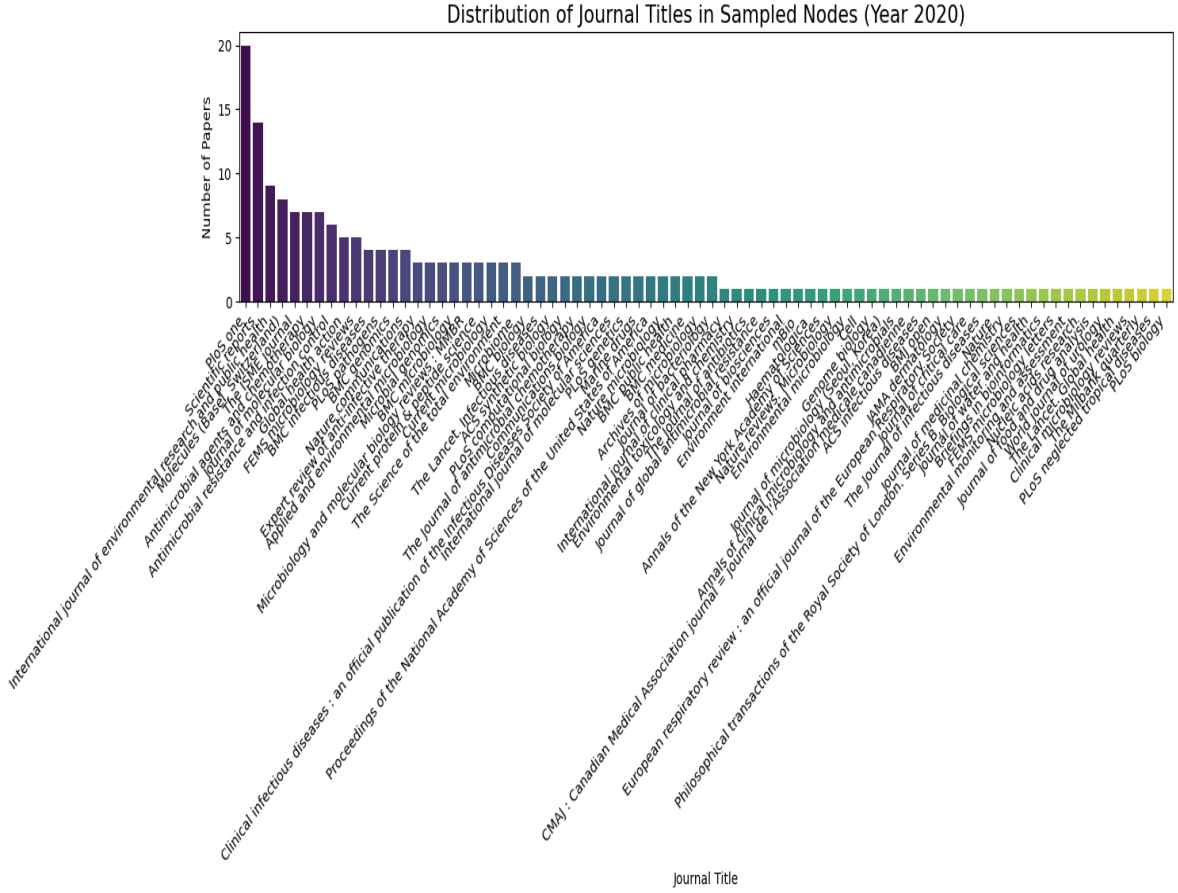

Figure 9: Counts of Journal titles in the sampled 200 nodes

4 Conclusion

Network analysis of various systems, such as citation networks, offers valuable insights into the underlying patterns and dynamics within a field. In this study, the citation network analysis revealed distinct characteristics of knowledge dissemination and research impact. Key metrics highlighted the presence of influential papers with high out-degree values, which serve as foundational works and survey articles that contribute to high in-degrees for other papers, driving the field forward.

The trend analysis over the years shows a steady growth in research activity from 1965, peaking around 1999, followed by a subsequent decline. This trajectory reflects the historical progression and evolving priorities within the field.

The examination of off-topic citations—papers not directly related to antimicrobial resistance (AMR) and drug resistance—further emphasizes the interdisciplinary nature of scientific research. The frequent incorporation of off-topic references demonstrates the significant influence of diverse scientific approaches and underscores the cross-disciplinary evolution of knowledge.

Additionally, the identification of highly cited journal titles and commonly used keywords provides insights into publication trends and research priorities within AMR studies. These patterns reveal the central themes and shifting focus areas that have shaped the field over time.

Overall, this network analysis highlights the interconnected, multidisciplinary nature of scientific research. It underscores how foundational works, emerging trends, and cross-disciplinary influences collectively contribute to the development and evolution of AMR studies and broader scientific inquiry.

References

1. Newman, M. *Networks* (Oxford university press, 2018).
2. McLaren, C. D. & Bruner, M. W. Citation network analysis. *International Review of Sport and Exercise Psychology* **15,** 179–198 (2022).
3. Que, X., Checconi, F., Petrini, F. & Gunnels, J. A. *Scalable community detection with the louvain algorithm* in *2015 IEEE international parallel and distributed processing symposium* (2015), 28–37.