Analysis of Microbiome Research using Co-Author Network

공저자 네트워크를 통한 마이크로바이옴 연구 분야 분석

by
Ha-Hyeon Kim

Under the supervision of
Professor Jung-Hye Roe, Yeong-Jae Seok, Ph. D. and Hong-Tak Lim, Ph. D.

August 2023

School of Biological Sciences
The Graduate School
Seoul National University

ABSTRACT

Analysis of Microbiome Research using Co-Author Network

Ha-Hyeon Kim
Biological Science
The Graduate School
Seoul National University

To identify the structure and causes of scientific research growth, we performed bibliographic analysis and network analysis. We selected the field of microbiome research, which has recently experienced rapid development and has a substantial research scale. Specifically, we choose the top 11 countries in terms of global microbiome research scale and analyzed the network of these countries from 2000 to 2021 to identify commonalities and differences in research network changes per country.

Through bibliographic analysis, we confirmed that the rapid growth in the microbiome field began in the early 2010s. The growth timing of most countries was behind that of the United States. Of interest, we observed a phenomenon of research scale reversal between the U.S. and China. By conducting bibliographic analysis, we were able to identify consistent growth patterns, enabling us to make predictions about the expansion of science networks.

To further investigate this approach, we constructed a Co-Author Network, which represents interconnections among scientists and plays a pivotal role in the scientific research network. By analyzing the quantity of network nodes that represent authors, we achieved the ability to forecast network growth. The growth of networks in the majority of countries closely aligned with the predictions derived

from bibliographic analysis.

Using various metrics from network theory, we examined the structure of the network. Initially, we looked at the Average Clustering Coefficient (ACC), which quantifies the cohesiveness among adjacent nodes. The scientific research network exhibited a high ACC value from its inception, gradually decreasing over time. This suggests that researchers within the network maintained strong connections throughout its growth, albeit slightly decreasing over time. Notably, we observed a faster decline in ACC within the Chinese network compared to other countries.

The Average Path Length (APL), which indicates network information efficiency, displayed a unique pattern. While an increase in APL is typically associated with a growing number of network nodes, most networks showed an S-shaped increase that eventually converged to a certain value. However, in certain countries, APL followed a linear increase and also converged. This finding highlights the previously unknown phenomenon of APL convergence during network creation and development. Additionally, our research demonstrated the applicability of network creation models such as the "Erdős-Rényi model" and the "Watts-Strogatz model" to real-world network cases. Based on our results, we predict that the microbiome research field will approximate a "Small World network," characterized by a high ACC and short APL, akin to the Watts-Strogatz model. Particularly during its early stages, the network closely approximates a Small World network and continues to do so as it grows. However, if the ACC gradually decreases, as observed in APL convergence at a certain distance, the continuous approximation to the Small World network may be compromised. The rapid decline of ACC, as seen in China, could be interpreted as an example of the swift randomization of research networks.

According to the "Erdős-Rényi model," a giant component, representing the largest connected component in the entire network, tends to emerge in sufficiently large and dense networks. Throughout the growth period of the 11 countries, all of them exhibited giant components. However, we also observed that the proportion of giant components did not directly correlate with changes in network size. This observation implies the differences in network structure development among countries.

Next, we examined the characteristics of key nodes in each network to elucidate the factors driving network growth and scientific progress. Specifically, we analyzed Betweenness centrality, which quantifies the number of times a node serves as a bridge between different researchers. The status of key nodes in each network showed dynamic changes as the network developed. Notably, the top nodes in the network did not maintain a consistent position in terms of connectivity over the analysis period, indicating a dynamic nature of scientific network development. In the cases of the United States and China, nodes related to technology emerged as rising stars within their respective networks and demonstrated connections with other country networks. This highlights the significant influence of technology among the factors driving network development.

Keywords: Co-Author Network, Microbiome research, Network analysis, High-throughput sequencing, Technological advancements, Collaborative patterns, Interdisciplinary collaborations

Student Number: 2008-20351

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBEREVIATIONS

APL	Average Path Length
ACC	Average Clustering Coefficient
TS	Topic sets
WoS	Web of Science

Chapter I. INTRODUCTION

I.1. Advancements in microbiome research

The microbiota refers to the collection of microorganisms, including bacteria, archaea, lower and higher eukaryotes, and viruses residing in a specific environment (Lederberg and McCray, 2001). Contrasting this, the term ‘microbiome’ paints a more comprehensive picture, encompassing not only the microorganisms but also their genomes, and the surrounding environmental conditions (Marchesi and Ravel, 2015). Despite the obscure origin of the term "microbiome" (Prescott, 2017), the investigation into microorganisms in defined environments can be traced back to the late 17th century, particularly in the context of humans, notably marked by Antonie van Leeuwenhoek’s innovative use of his newly-developed microscopes (“Milestones in Human Microbiota Research,” n.d.).

The field of microbiome research has witnessed a remarkable surge in recent years, attributed largely to technological breakthroughs and the subsequent cost reduction of analysis methodologies. The introduction of high-throughput sequencing technologies has revolutionized the capacity to investigate the microbial communities in various environments, thereby aiding in the understanding of their composition and dynamics. Thus, in turn, it has facilitated an exponential upsurge in the generation of metagenomic data over the past decade (Waldor et al., 2015).

Within the realm of microbiome research, the field can be broadly categorized into the following areas (Cullen et al., 2020):

1. Host-microbe interactions: This area primarily focuses on the complex interrelationships between microorganisms and the human host, including their impact on health, disease, and overall well-being.

2. Microbial evolution and ecology: This area is dedicated to elucidating the evolutionary and ecological aspects of microorganisms, emphasizing the impact of environmental factors and microbe-microbe interactions on microbial communities.

3. Analytical and mathematical approaches in microbiome research: This domain involves the application of various analytical and mathematical strategies to process and interpret microbiome data, enabling in-depth analysis and exploration of complex microbial communities.

4. Bioengineering solutions based on microbial composition and interactions: Leveraging knowledge of microbial composition and interactions, researchers are investigating potential engineering solutions for various applications, including bioremediation, bioenergy, and biotechnology.

5. Interventional strategies and engineered microbiota: This area explores the feasibility of strategically modifying microbial composition to yield specific outcomes, including the development of engineered microbiota for therapeutic interventions and other applications.

The interdependence between analytical and mathematical methods (Brooks, 1994) in microbiome research and the emergence of high-throughput sequencing technology showcases the symbiotic relationship between science and technology. These advancements have revolutionized our understanding of the microbiome, setting a strong foundation for further advancements in this rapidly evolving field.

I.2. Social Network Analysis

Social network analysis focuses on the study of relational data, which includes connections, ties, group affiliations, meetings, and other interactions amongst individuals. Unlike attribute data that examine individual characteristics, relational data cannot be reduced to individual attributes and instead capture the characteristics of the entire system. Network analysis involves analyzing and interpreting these relational data, aiming to comprehend the underpinning of the structure of social action (Scott, 2012).

Social science data can be broadly classified into three categories: attribute data, relational data, and ideational data. Attribute data pertain to the attributes, opinions, and behaviors of individuals and are analyzed through variable analysis, which unravel correlations between variables and outcomes. On the other hand, relational data focuses on capturing the interactions, relationships, and group memberships among individuals, utilizing network analysis techniques for analysis. Ideational data describe meanings, motives, definitions, and classifications which hold a pivotal role in social science, albeit technological solutions for managing such data are still evolving.

The significance of studying of relational data lies in its capacity to elucidate the fabric of social structures, inherently rooted in human relationships. Therefore, exploring social structures often requires obtaining and analyzing relational data, which reveals the patterns and dynamics of social interactions.

Historically, the analysis of human behavior and psychology was primarily conducted through subjective introspection, analyzing the human mind as a constituent element. However, the advent of Gestalt theory (Smith, 1988) challenged this approach by emphasizing the importance of groups and social environments in shaping individual experiences. This paradigm shift fueled a growing interest in comprehending the structural intricacies of groups and societies.

To visually represent the characteristics of social structures, Jacob Moreno and

Helen Jennings introduced sociograms in 1934. Sociograms depicted individuals with popular appeal and leadership qualities as "stars" positioned at the center of their relationships in the social structure diagram (Scott, 2012).

Building upon these developments, Konig advocated for the use of graph theory in 1936 to simplify complex relationships and represent the entire social structure. According to this view, the repetition of such relationships forms the social structure.(Scott, 2012)

Subsequently, the approach outlined by Gestalt theory to encapsulate the influence of group dynamics and social environments on individuals matured further. By the 1940s, it became possible to represent the entire social structure by simplifying complex relationships. This method of representation facilitated researchers like Anatol Rapoport in the 1960s to introduce models for analyzing infectious diseases based on social interrelationships (Scott, 2012). Additionally, to describe the properties of networks, several metrics have been devised and applied by physicists and mathematicians (Stanley, 1967). One key aspect of network analysis is the examination of different network models that capture distinct characteristics of real-world networks. Random networks, small-world networks, and scale-free networks are three prominent types that have been extensively studied. Each of these network models offers unique features and sheds light on different aspects of network behavior and dynamics.

There are several classes of networks, each having unique characteristics and behaviors.

Regular Networks: A regular network is a network where each node has the same number of connections. An example of this would be a lattice or a grid where each point is connected to its nearest neighbors. The key characteristic of regular networks is their regularity and predictability. They lack complexity and diversity, and their topology is usually homogeneous.

Random Networks: Random networks are generated through a process where

connections between nodes are established randomly. They were first studied by mathematicians Erdős and Rényi in the late 1950s. In a random network, there's an equal probability of any pair of nodes being connected. These networks are highly unpredictable and their properties can vary drastically, depending on the randomness.

Small-World Networks: Small-world networks are a class of networks where most nodes can be reached from every other node in a small number of steps. These networks are characterized by high clustering coefficients and short average path lengths. A real-world example of this is the concept of 'six degrees of separation', which suggests that any two people on Earth are, on average, separated by only six acquaintance links.

Scale-Free Networks: Scale-free networks are networks whose degree distribution follows a power law, at least asymptotically. This means that there are few nodes with many connections (hubs), and many nodes with few connections. The structure of these networks is influenced by two major mechanisms: growth and preferential attachment, where new nodes are more likely to connect to nodes that already have many connections. The World Wide Web and social networks are common examples of scale-free networks.

Non-Scale-Free Networks: Non-scale-free networks are those networks that do not follow the power law degree distribution. The connections in these networks are not dominated by a few highly connected nodes. They tend to be more homogeneous in their degree distribution, meaning there isn't a significant difference in the number of connections each node has.

I.3. Purpose of Research

The purpose of this research is to bridge the gap in quantitative studies that accurately depict the evolution of the microbiome field over the past two decades. The primary objective is to provide an in-depth, data-driven analysis of the transformation of field during this era.

To achieve this, the study focuses on analyzing the network of authors who have contributed to the corpus of research papers in the Microbiome and Microbiota fields. Specifically, the analysis focuses on Co-Author Network analysis examination, with a particular emphasis on the shifts in co-author networks over time. This innovative approach offers a novel and previously uncharted method of interpreting the dynamics within the field. The analytical process was governed by following flowchart, which served as a roadmap for the data collection and analysis (Fig. I.1).

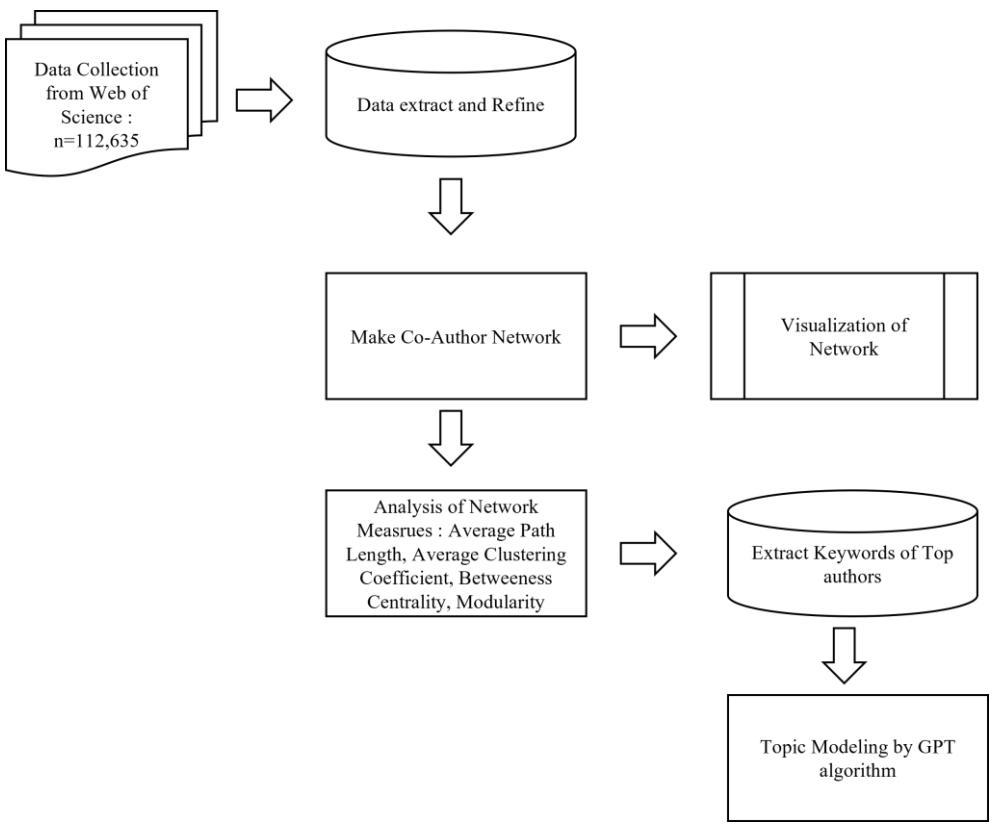

Figure I.1. Flow Chart of Network Analysis and Topic Modeling

The data was collected from Web of Science, which provides comprehensive information including author names, affiliations, abstracts, and keywords. The boxes in the figures represent the findings of each research paper. And the cylinders represent the progress of the analysis.

By leveraging analytical measures of network methodology, this research

strives to offer a more quantitative depiction of the changes and trends within the microbiome field. Additionally, topic modeling techniques will be applied to supplement the analysis and visualize the evolving of research themes.

It is important to acknowledge that the advancement of life sciences is influenced by various factors, including societal demand (Dolgin, 2017; Kemp, 2015), social relationships among researchers (de Siracusa et al., 2020), and technological advancements (Stella and Rem, 2017). By exploring the interplay between these factors and the expansion of microbiome research, this study aspires to illuminate the driving forces behind its growth.

The interdependence between science and technology is a fundamental aspect that warrants consideration. It is widely acknowledged that advancements in either domain profoundly impact the evolution of the other. The relationship between science and technology is symbiotic, with each driving and benefiting from the other's advancements. Technology propels scientific inquiry forward. Technological advancements provide scientists with innovative tools, instruments, and methodologies to conduct research more efficiently and accurately. These cutting-edge technologies enable scientists to explore new frontiers, collect vast amounts of data, and analyze complex phenomena in ways that were previously unimaginable (Brooks, 1994).

In South Korea, previous studies have elucidated the disparities between domestic microbiome research and international efforts (Park et al., 2014). It has been noted that Korean researchers have not been at the forefront compared to researchers from other countries, particularly in terms of patents registrations. Furthermore, the proportion of patents in the fields of hygiene and food has been significantly higher in Korea compared to other countries. Additionally,a downturn in research activities among certain researchers has been observed. These discrepancies in societal demand, interpersonal relationships amongst researchers, and technological advancements may explain the differences in microbiome research between Korea and other countries.

It is worth noting that researchers often tend to study subjects that are accessible and convenient for their research purposes (Daru et al., 2018). Such tendency can lead to research concentrations or trends in specific fields following the emergence of new scientific paradigms or technological advancements. Understanding these dynamics in the context of microbiome research will provide invaluable insights into its growth trajectory.

In summary, this research endeavors to foster a comprehensive understanding of the growth of the microbiome field, pinpoint pivotal factors driving its evolution, including technological advancements, and delve into the underlying processes via the application of Network Analysis and other relevant methodologies.

Chapter II. MATERIALS AND METHODS

II.1. Data collection

We conducted our research on a dataset comprising 112,635 papers extracted from the Web of Science citation index database provided by Clarivate Analytics. The papers were selected based on the search terms "Microbiome" or "Microbiota" (Table 1.).

Data analysis was performed using Python v.3.6 (Guido, 1995), leveraging the capabilities of the Pandas (McKinney, 2010) and Numpy packages (Harris et al., 2020). The Python programming language is widely recognized for its flexibility and extensive range of scientific computing libraries.

To refine and analyze the data, we utilized BASH (GNU, 2007), a command-line shell and scripting language commonly used in data processing and manipulation tasks. The BASH environment facilitated necessary data refinement and preparation for subsequent analysis.

The computations were carried out on a system equipped with an Intel® Core™ i5-6600 CPU @ 3.30GHz, 16.0GB RAM, and a GeForce GTX 1060 3GB graphics card. This computing setup provided the necessary resources to perform the calculations efficiently and effectively.

By employing these tools and resources, we were able to conduct a thorough analysis of the dataset and derive meaningful insights into the microbiome field.

Table II.1.1. Search terms for analysis

TS	**Microbiome or Microbiota**
Year	2000-2021
Type of Article	Article only

'TS' stands for Topic Set and it represents the topics derived from the classification in the Web of Science.

II.2. Co-Author Networks (CANs)

To construct the co-author network, we employed the "AU" field of the Web of Science database, which contains author information. Each author listed in the "AU" field was treated as a distinct node within the network. Connections were established between authors who co-authored papers together, capturing their collaborative relationships by Tethne. Additionally, authors with the same name appearing in different papers were linked to account for their shared identity. This information was organized into a matrix representation (Peirson, 2016).

The resulting matrix was visualized using the Gephi software (Bastian et al., 2009), allowing for a comprehensive visual exploration of the co-author network. In addition to visualization, various network measures were computed to gain insights into the network's characteristics. These measures included average path length, betweenness centrality, and clustering coefficient, which provide quantitative assessments of network connectivity and centrality of individual authors.

By utilizing these techniques and tools, we were able to analyze and visualize the co-author network, uncovering patterns of collaboration and identifying key authors within the microbiome research field.

II.3. Measures of network

II.3.1. Average Path Length

The average path length is a fundamental measure of network efficiency (Newman, 2004), indicating how easily and quickly information or mass can flow within a network. It is a critical factor in various types of networks, including the Internet, metabolic networks, and power grids. A shorter average path length in these networks enables faster information transfer, reduces costs, and minimizes losses.

In network theory, the concept of a "small world" describes the remarkable interconnectedness observed in many real-world networks, despite their large size. This term was coined by sociologist Stanley Milgram (Stanley, 1967), who conducted the famous "six degrees of separation" experiment to study social network connectivity (Latapy, 2008). The small-world phenomenon suggests that most real networks exhibit a relatively short average path length, enabling efficient communication and rapid dissemination of information.

Mathematically, the average path length of a connected graph G is calculated as the average distance between pairs of vertices, known as the PathLengths. The average path length L(G) is defined as the sum of all pairwise distances divided by the total number of possible vertex pairs in the graph.

$$d = Distance$$

$$L(G) = \frac{1}{n(n-1)} \sum_{u,v \in V} d(u,v)$$

$$where\ d(u,v) = 0 \quad if \quad u = v$$

The average path length typically depends on the size of the network, but it follows a logarithmic relationship with the number of nodes (n) according to small-world network theory. This means that as the network grows, the average path length increases proportionally to the logarithm of the number of nodes.

Studying the average path length provides valuable insights into the structure and efficiency of networks, aiding our understanding of their information flow dynamics and overall performance.

II.3.2. Average Clustering Coefficients

The clustering coefficient is a measure that quantifies the tendency of nodes in a graph to form clusters or tightly-knit groups. It reflects the density of connections among the neighbors of a node and provides insights into the degree of clustering within the network (Latapy, 2008).

Mathematically, the clustering coefficient of a vertex v in a graph G is calculated as the ratio of the number of pairs of adjacent neighbors to the number of pairs of all possible neighbors of v. The clustering coefficient of the entire graph G, denoted as CC(G), is obtained by taking the average of the local clustering coefficients over all nodes in the network.

$$v = vertex\ in\ a\ graph\ G$$

$$cc(v) = \frac{number\ of\ pairs\ of\ adjacent\ neighbors}{number\ of\ pairs\ of\ neighbors}$$

$$CC(G) = clustering\ coefficeint\ of\ a\ graph\ G$$

$$CC(G) = \frac{1}{|V|}\sum_{v \in V} CC(v)$$

$$0 \leq CC(G) \leq 1$$

$$CC(G) \approx 1 \rightarrow Graph\ is\ higly\ clustered$$

$$CC(G) \approx 0 \rightarrow Graph\ is\ not\ highly\ clusterd$$

The clustering coefficient values range between 0 and 1, where a value close to 1 indicates a highly clustered graph, implying that nodes tend to be strongly connected within their local neighborhoods. Conversely, a value close to 0 suggests a graph with lower clustering, indicating a more dispersed or loosely connected network structure.

The average clustering coefficient, CC(G), provides an overall measure of the clustering tendency within the entire network. It indicates the average level of clustering observed across all nodes in the graph. On the other hand, the local clustering coefficient, cc(v), measures the clustering tendency of an individual node

by examining the connectedness of its immediate neighbors.

The clustering coefficient is a valuable measure for understanding the structural properties of networks and gaining insights into their functional behavior. For instance, in social networks, a high clustering coefficient often indicates the presence of tightly-knit communities or groups, highlighting the cohesive nature of social relationships. In transportation networks, a low clustering coefficient suggests the existence of multiple alternative routes between locations, promoting efficient transportation flow.

Analyzing the clustering coefficient can help reveal the organization and connectivity patterns within a network, shedding light on its resilience to disruptions, information diffusion processes, and overall system dynamics.

II.3.3. Modularity

Modularity is a fundamental concept in network analysis that measures the division of a network into distinct communities or modules. It quantifies the degree to which nodes within the same module are more densely connected to each other compared to nodes in different modules.

This concept finds broad application in various domains, including social networks, biological networks, and technological networks. It provides insights into the organization and structure of complex systems. By identifying modules within a network, researchers can gain valuable insights into the relationships between nodes and the formation of clusters or communities (Blondel et al., 2008).

The Louvain method is a commonly used algorithm for calculating modularity. It aims to identify modular structures by optimizing the modularity score of a network. The algorithm consists of two main phases: the optimization phase and the aggregation phase.

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

$$Aij = edge\ weight\ between\ nodes\ i\ and\ j$$

$$k_i, k_j = sum\ of\ the\ weight\ of\ the\ edges\ attaced\ to\ node\ i\ and\ j$$

$$m = sum\ of\ all\ of\ the\ edge\ weights\ in\ the\ graph$$

$$c_i, c_j = communities\ of\ the\ nodes$$

$$\delta = Kronecker\ delta\ function\ (\delta(x, y) = 1\ if\ x = y, 0\ otherwise)$$

Modularity of a community c can be calculated:

$$Q_c = \frac{\sum in}{2m} - (\frac{\sum tot}{2m})^2$$

$$\sum in = sum\ of\ edge\ weights\ between\ nodes\ within\ the\ community\ c$$

$$\sum out = sum\ of\ all\ edge\ weight\ for\ nodes\ within\ the\ community$$

In the optimization phase, the algorithm iteratively assigns nodes to communities in a greedy manner to maximize the increase in modularity. It starts with each node in its own community and iteratively evaluates the modularity gain by moving nodes to neighboring communities. This process continues until no further improvement in modularity can be achieved.

After the optimization phase, the algorithm enters the aggregation phase, where the communities obtained in the previous phase are treated as individual nodes. The network is then represented as a new graph, where each community becomes a node. The edges between the new nodes are weighted based on the sum of the edge weights between nodes in the original network that belong to different communities. This step reduces the complexity of the network while preserving its modular structure.

The optimization and aggregation phases are iteratively repeated until a maximum modularity value is achieved, or a stopping criterion is met. At the end of the algorithm, the resulting community structure represents the detected modules in the network.

Modularity can be calculated at both the community level and the network level. The modularity of a community is calculated based on the sum of the edge weights within the community compared to the expected sum of edge weights if connections were randomly distributed. The network-level modularity is the sum of the modularity values of all communities.

$$\Delta Q = \left[\frac{\sum in + 2k_{i,in}}{2m} - (\frac{\sum tot + k_i}{2m})^2\right] - \left[\frac{\sum in}{2m} - (\frac{\sum tot}{2m})^2 - (\frac{k_i}{2m})^2\right]$$

$\sum tot$

$= sum\ of\ all\ the\ weights\ of\ the\ links\ to\ nodes\ in\ the\ community\ is\ moving\ into$

$k_i = weighted\ drgree\ of\ i$

$k_{i,in}$

$= sum\ of\ the\ weights\ of\ the\ links\ between\ i\ and\ other\ nodes\ in\ the\ community$

$$m = sum\ of\ the\ weight\ of\ all\ links\ in\ the\ network$$

Modularity has diverse applications in social network analysis, recommendation systems, gene expression analysis, and identifying functional modules in biological systems. It provides valuable insights into the underlying organization and dynamics of complex systems.

II.3.4. Betweenness Centrality

In network analysis, betweenness centrality is a measure that captures the importance of a vertex within a graph based on the concept of shortest paths. Betweenness centrality has long been recognized as a crucial aspect of centrality, with Freeman providing the first formal definition (Freeman, 1977).

For any pair of vertices in a connected graph, there exists at least one shortest path between them. This path minimizes either the number of edges in unweighted graphs or the sum of the weights of edges in weighted graphs. The betweenness centrality of a vertex is determined by the number of shortest paths that pass through it (Brandes, 2001).

Mathematically, the betweenness centrality of a vertex v can be calculated by summing up the fraction of shortest paths passing through v for all pairs of vertices (s, t), excluding cases where v is an endpoint:

$$\sigma_{st} = total\ number\ of\ shrtest\ paths\ from\ node\ \boldsymbol{s}\ to\ node\ \boldsymbol{t}$$

$$\sigma_{st}(v) = number\ of\ paths\ through\ v\ (not\ where\ v\ is\ an\ end\ point)$$

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

Betweenness centrality finds widespread application in network theory as it quantifies the extent to which a node acts as a bridge between other nodes in the network. In a telecommunications network, for example, a node with high betweenness centrality would exert greater control over the network's flow of information as more paths pass through it.

II.4. Topic Modeling

To facilitate topic modeling, the following steps were undertaken in this

research:

Model Architecture: GPT (Generative Pre-trained Transformer) (OpenAI, 2023) and BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) were utilized as powerful language models. GPT employs a unidirectional transformer architecture and predicts the next word in a sequence based on the context of previous words. On the other hand, BERT is a bidirectional model that captures contextual information from both left and right contexts, enabling a deeper understanding of word meaning within a sentence.

Training Objective: GPT and BERT were pre-trained using unsupervised learning on extensive corpora. GPT focuses on generating coherent and contextually appropriate text, while BERT aims to generate contextualized word representations by comprehending the relationships between words in a sentence. Both models serve different purposes, with GPT emphasizing text generation and BERT emphasizing contextual understanding and classification tasks.

Modularity Calculation: Modularity was computed for each network in the dataset. The networks were divided into communities or modules using the Louvain algorithm, which optimizes the modularity score by iteratively partitioning the network. Higher modularity scores indicate a stronger division into distinct modules. The top 8 networks with the highest modularity scores were selected for further analysis.

Identification of Key Authors: Within the selected networks, the author with the highest betweenness centrality was identified. Betweenness centrality measures the extent to which an author serves as a bridge between different communities in the network. By selecting the author with the highest betweenness centrality, the research aimed to identify individuals playing a crucial role in connecting different research topics.

Collection of Keywords: The Web of Science (WoS) dataset was utilized to gather keywords associated with the publications of the selected authors. This

involved retrieving publications attributed to the identified authors and extracting the associated keywords. The objective was to capture the primary research themes and topics explored by these authors.

Topic Modeling with GPT Algorithm: The collected keywords were employed as input for topic modeling using the GPT 3.5 algorithm. Topic modeling is a statistical technique that uncovers underlying themes or topics within a collection of documents. By leveraging the GPT 3.5 algorithm, latent topics were automatically extracted from the keyword dataset, enabling the exploration of the main research themes present in the selected networks.

By following these steps, this research aimed to reveal the modular structure of the networks based on modularity scores, identify influential authors based on betweenness centrality, gather relevant keywords from their publications, and ultimately employ the GPT algorithm for topic modeling to uncover the primary research topics within the networks.

Chapter III. RESULTS

III.1. The Growth of microbiome research

The field of Microbiome research has experienced significant and rapid growth over the past two decades. In 2021 alone, an impressive tally of over 110,000 research papers was published in this field, involving a staggering number of more than 360,000 researchers as authors (Fig. III.1).

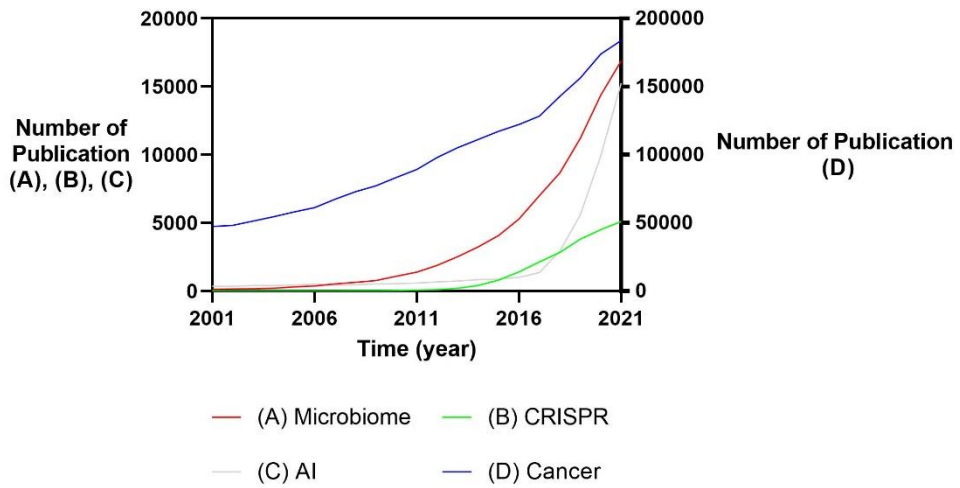

Figure III.1.1. Comparison of annual publication numbers by research topic

Left: The axis representing the number of publications for (A), (B), and (C).
Right: The axis representing the number of publications for (D).

For a comparative viewpoint, we examined publication trends in other prominent fields, including Cancer research, AI research, CRISPR research, and Microbiome research. Our analysis revealed that the Microbiome field underwent a remarkable surge in the early 2010s. While the Cancer research exhibited a larger volume of publications in absolute terms, its growth was consistent rather than rapid swift. In contrast, the CRISPR field experienced a concentrated period of rapid growth at a particular juncture. Interestingly, Microbiome research exhibited an exponential increase in publication output, exceeding the CRISPR field by more than

threefold. This growth trajectory closely resembled the expansion observed in the field of artificial intelligence (AI).

These findings highlight the dynamic and thriving nature of the microbiome research, underscoring its escalating importance and applicability in scientific research. The exponential surge in publications indicates the growing interest and acknowledgment of the microbiome's significance across various disciplines, including health, ecology, and biotechnology. As the field continues to expand, it holds considerable potential for further advancements and discoveries that could enhance our understanding of the microbiome's role in human health and the environment.

The average number of authors per journal article has exhibited a steady and upward trend, increasing from 3.6 to 5.9 over time (Fig. III.1.2). This growth signifies a notable shift in the collaborative dynamics of microbiome research. Previous studies have reported an average of 3.754 authors in Medline journals between 1995 and 1999, a figure closely matched by the 3.6 observed in the microbiome field in the year 2000 (Newman, 2001).

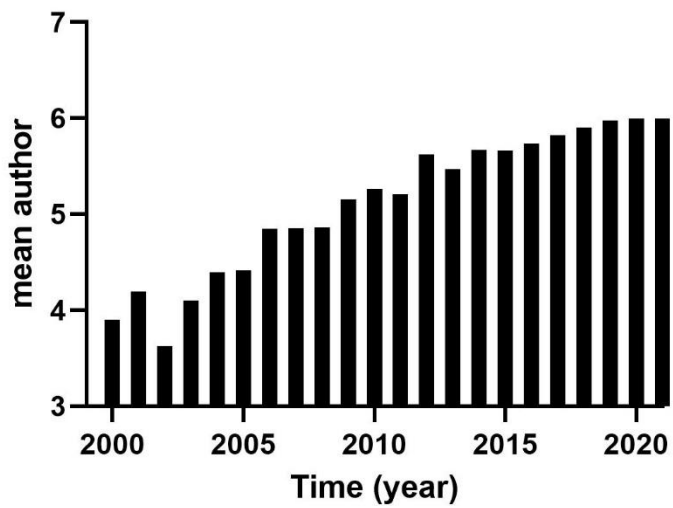

Figure III.1.2. Mean number of authors per paper

The mean author refers to the average number of authors per article. It is calculated by dividing the total number of authors in a given year by the total number of articles published in that year.

The increasing average number of authors per journal article is indicative of the evolving nature of microbiome research. It reflects a transition towards heightened interdisciplinary collaboration and the inclusion of larger research teams with diverse areas of expertise. This tendency suggests that tackling complex questions and exploring the intricacies of the microbiome requires a collective effort and the integration of various scientific perspectives.

Overall, the rise in the number of authors per article highlights the growing complexity and multidimensionality of microbiome research. As the field expands, researchers from different disciplines, including microbiology, immunology, genomics, bioinformatics, ecology, and clinical medicine, are pooling resources to investigate the complex interplay between microorganisms and their hosts. This interdisciplinary collaboration fosters a more comprehensive understanding of the microbiome and its impact on human health and the environment.

Given the evolving nature of microbiome research, it is essential to maintain a continuous monitoring and analysis of these changes. This ongoing examination will provide valuable insights into the collaborative dynamics, emerging trends, and interdisciplinary nature of the field. Such insights will enable researchers to adapt their strategies and foster effective collaborations, thereby facilitating further advancements in the field of microbiome science.

III.2. Comparison of the microbiome field by country

Upon categorizing the data from the Web of Science (WoS) by nation, it was evident that the United States and China have contributed substantially to microbiome research. Both countries are responsible for over 19,000 publications, exhibiting a greater publication output compared to other nations (Fig. III.2.1).

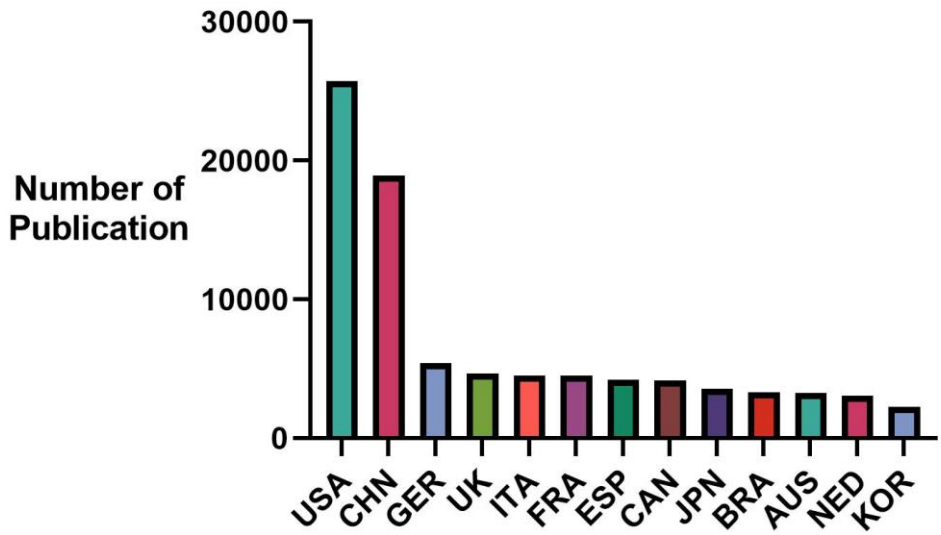

Figure III.2.1. Comparison of the number of microbiome papers by country

The number of publications in the figure represents the total number of publications for each country from 2000 to 2021.

When comparing the total number of publications between Germany and South Korea, there is a notable difference of approximately 2.5 times. Germany has a total publication count of 5,438, while South Korea has 2,289 publications. This discrepancy highlights the varying levels of research output and activity in the two countries within the [insert specific field or domain]. Understanding these differences in publication numbers can provide insights into the research landscape and potential areas for collaboration or further investigation.

Figure 5 illustrates the changes in the number of research publications by country. The United States has experienced a significant increase in microbiome research publications since 2010, demonstrating a consistent growth pattern. Conversely, a rapid surge in publications is observed around 2015 in China. (Fig III.2.2.A) Other Nation's growth trajectory appears similar to that of China, albeit lagging by approximately five years compared to the United States and China. (Fig III.2.2.B)

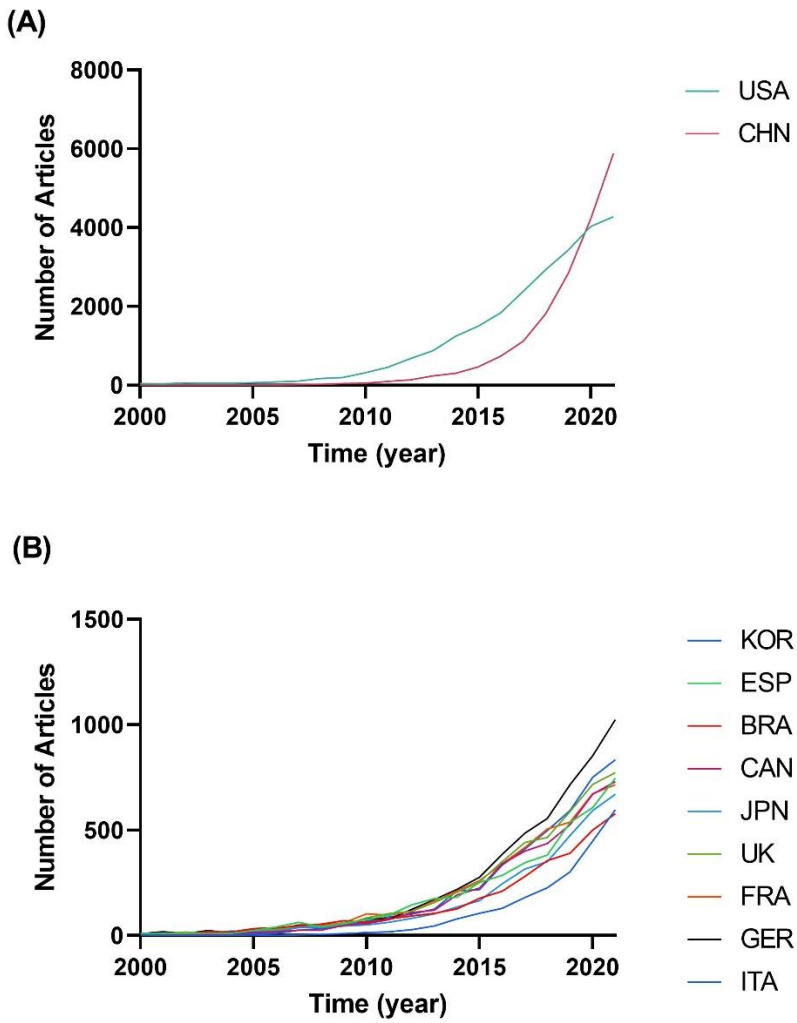

Figure III.2.2. Changes in the Number of Research Publication by Country

(A) The countries with annual publications exceeding 2000 in 2021 are the United States and China.

(B) The countries with annual publications below 2000 in 2021 are various other countries.

Despite South Korea sharing a similar growth pattern in terms of timing, it has a smaller absolute number of research publications than China. This observation suggests that while South Korea is narrowing the gap and showing a positive trend in microbiome research, it still has some ground to cover in terms of publication output.

These findings highlight the global distribution of microbiome research and the contributions made by different countries. The United States and China, being leaders in this field, have demonstrated significant influence in terms of research output. While South Korea is demonstrating promising growth, there is still scope for additional development and amplification of its research productivity.

By analyzing the changes in research publications by country, this study provides insights into the global landscape of microbiome research and the disparate contributions from different nations.

III.3. Analysis of network topology

In our analysis of network topology, we have placed significant emphasis on robust metrics such as average path length, clustering coefficient, and degree distribution to gain a profound understanding of networks and develop precise models (Blondel et al., 2008; Lambiotte et al., 2014; Latapy, 2008).

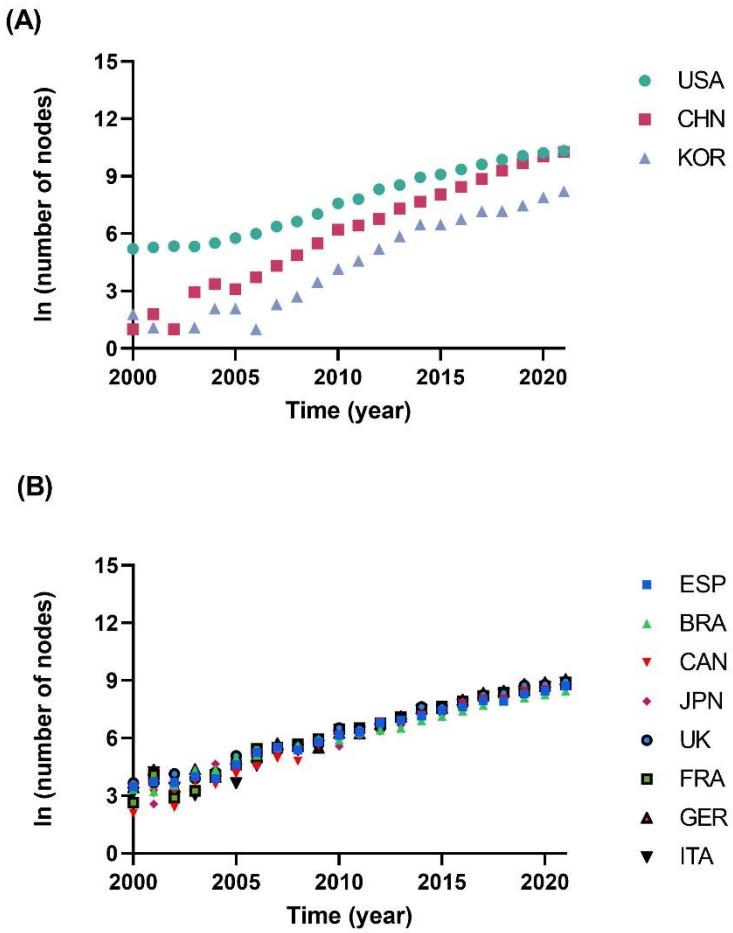

Figure III.3.1. Number of nodes of microbiome research Co-Author Networks in various countries

(A) The changes in the number of nodes for the United States, China, and South Korea are represented by a logarithmic function.

(B) The changes in the number of nodes for Spain, Brazil, Canada, Japan, the United

Kingdom, France, Germany, and Italy are represented by a logarithmic function.

In order to compare the size of networks, the number of nodes was examined for each country (Fig. III.3.1). According to (A), the United States starts at approximately 5.204007 in the year 2000 and reaches 10.3296 in 2021. China, on the other hand, starts at nearly 1 in 2000 and reaches 10.26955 in 2021. Looking at the other countries in (B), it can be observed that they generally reach the high 8s in 2021.

From this analysis, it is evident that the United States and China exhibit different patterns compared to the other countries. However, the differences among the other countries are relatively small, as previously mentioned. As depicted in Figure III.2.2, the number of nodes is known to be proportional to the natural logarithm (ln) of the Average Path Length (APL), which will be further explained below.

The changes in APL exhibit distinct patterns across various countries (Fig III.3.2). In the United States, a growth in APL was noted starting from 2008.

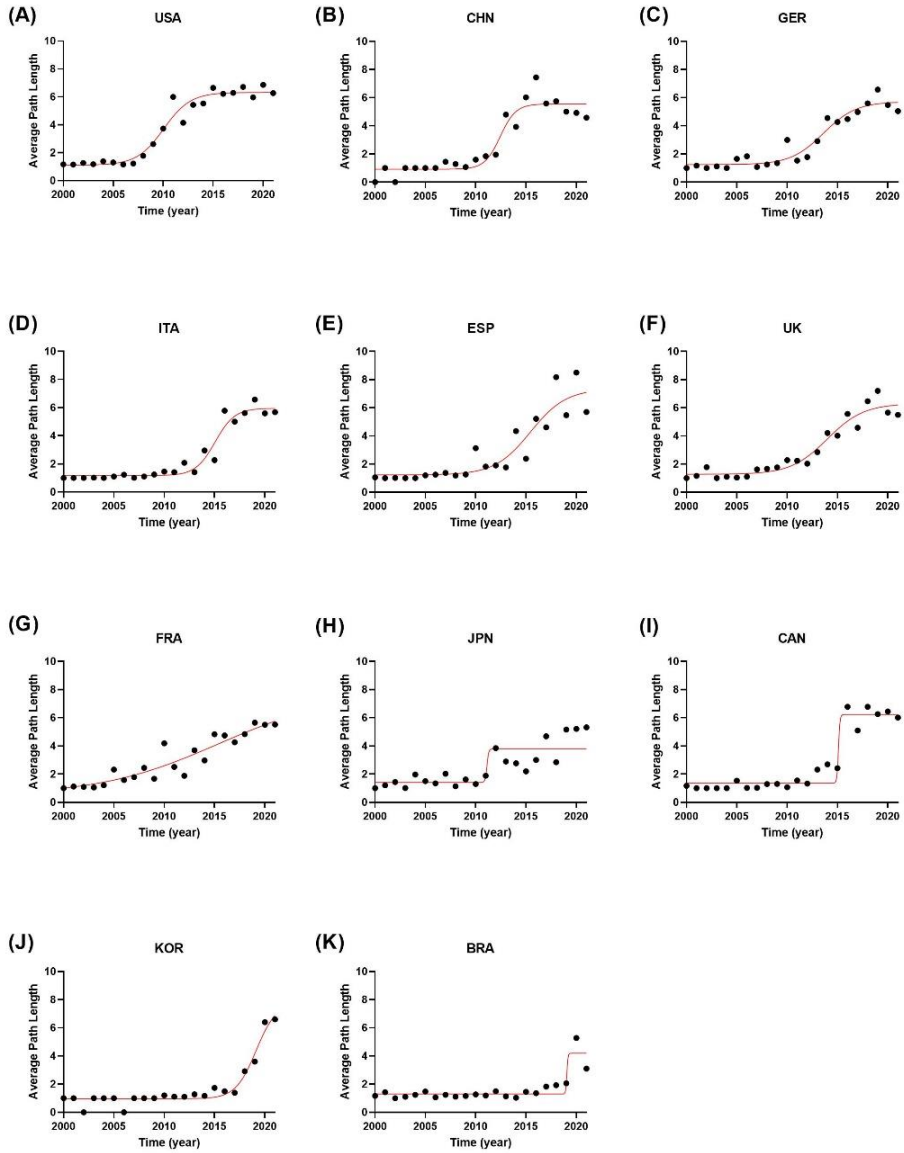

Figure III.3.2. Average Path Length of microbiome research Co-Author Network in various countries

The black dots represent the actual Average Path Length (APL) for each country, while the red solid line represents the trend line.

China experienced an increase in APL around 2012, subsequently decreasing to around 4 in 2021. South Korea, on the other hand, has shown a consistent growth trend in APL.

Previous studies have indicated that Medline journals in the field of biology exhibit a lower APL of 4.6 compared to 7.73 in mathematical research and 5.9 in physics research, indicating a significant "small network" characteristic (Newman, 2001).

The small world phenomenon, characterized by short average path lengths (APL) and high clustering coefficients, has been observed in various systems, including social networks, biological networks, and technological networks including the internet. This phenomenon stems from a balance between local connections fostering clustering and global connections enabling efficient long-range communication, contributing to the resilience and robustness of small world networks.

The rapid increase in APL of microbiome research network could be attributed to the formation of a large scientific collaboration network as the number of nodes (coauthors) grows, reflecting the characteristics of a "small world" network. Initially, microbiome researchers were loosely connected, resulting in a smaller APL. However, with the formation of researcher connections, a network with a small world structure emerged and grew.

The differences in APL, relative to the number of nodes, observed in Korea during the 2020s (Fig III.3.2.J), indicate significant differeneces compared to other countries. The APL in the United States remains around 6, while in China, it initially increased to 8 but gradually decreased over time.

These findings highlight the diverse characteristics of APL, including the timing of network formation, the size of APL, and the patterns of change, which differ among countries. These differences can be attributed not only to the number of nodes but also to variations in research fields across countries.

Countries such as France, along with other European countries and Japan,

exhibit a gradual increase in Average Path Length (APL) as shown in (Fig. III.3.2. C~H). On the other hand, Canada, South Korea, and Brazil demonstrate a significant increase in APL during the 2010s (I~K). Despite the similarity in the changes in the number of nodes as shown in Figure 6, these differences highlight the fact that factors influencing the network vary across countries, extending beyond the number of nodes.

These variations in the APL trends suggest that there are other factors at play, shaping the network dynamics within each country. It could be attributed to differences in research collaborations, funding, institutional structures, or scientific cultures, among other variables. Understanding these country-specific factors is crucial for a comprehensive analysis of the microbiome research landscape and its development across different nations.

The average clustering coefficients of all countries are high, indicating that researchers within each country efficiently form clusters in their respective networks. This elevated average clustering coefficient reflects the presence of tightly knit groups or research communities within each country's network.

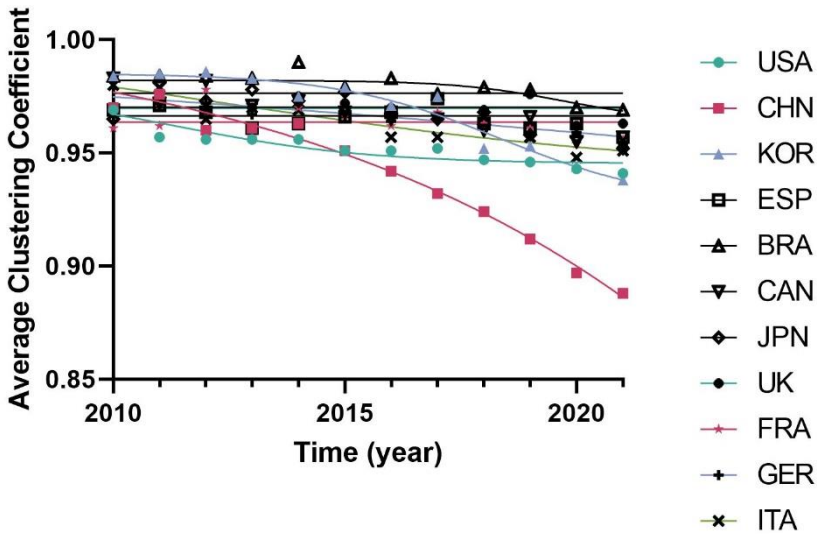

Figure III.3.3. Average Clustering Coefficients of microbiome research Co-Author Network in various countries

The graph represents the values of Average Clustering Coefficient (ACC) from 2010 to 2021. The maximum value of ACC is 1.

Interestingly, the average clustering coefficient in China shows a decrease compared to the other countries (Fig III.3.3). This indicates that as the overall network in China grows, the formation of clusters among researchers decreases on a global scale. In other words, the network in China is becoming more interconnected with researchers from different clusters collaborating, resulting in a decrease in the average clustering coefficient.

In contrast, the United States and South Korea maintain a high average clustering coefficient, indicating that researchers within these countries continue to form tightly knit clusters or communities in their networks. These findings underscore the

evolving dynamics of collaboration networks in the microbiome field, with China exhibiting a trend towards enhanced interconnectivity among researchers from different clusters.

III.4. Visualizations of Co-Author Networks

To gain a deeper understanding of the variations in the development of the microbiome field across different countries, we will closely examine the detailed communities within the networks. By analyzing the specific communities that have formed within each country's network, we aim to uncover the factors contributing to the observed differences.

By conducting a detailed analysis of the communities, we investigated the unique research themes, collaborations, and knowledge exchange patterns within each community. This will us to identify key drivers that have shaped the development of the microbiome field in each country.

Through this detailed examination of the network's communities, we aim to elucidate the specific factors and mechanisms that have influenced the growth and development of the microbiome field in different countries. By uncovering these insights, we can contribute to a comprehensive understanding of the field's evolution and provide a basis for future research and policy considerations.

Due to the difficulty of representing all the information in this research, we have chosen to focus on average path length using Figure III.3.2. We selected three specific time points, categorized as early, mid, and late stages, for each country. Modularity was then calculated for each network using a modularity resolution of 1.0.(Lambiotte et al., 2014) (Table III.4.1)

Table III.4.1. Time Point of Network Phase

	APL50	Earyl	Mid	Late
USA	2010	2006-2008	2009-2011	2012-2014
JPN	2011	2007-2019	2010-2012	2013-2015
CHN	2012	2008-2010	2011-2013	2014-2016
UK	2014	2010-2012	2013-2015	2016-2018
GER	2014	2010-2012	2013-2015	2016-2018
ESP	2015	2011-2013	2014-2016	2017-2019
CAN	2015	2011-2013	2014-2016	2017-2019
FRA	2015	2011-2013	2014-2016	2017-2019
ITA	2015	2011-2013	2014-2016	2017-2019
BRA	2019	2015-2017	2018-2020	n.d.
KOR	2019	2015-2017	2018-2020	n.d.

In Figure III.4.1.we observe the variations in the changes of the Giant Component across different countries. Generally, the Giant Component increases over time for most countries, indicating the growth of connectedness within the networks. However, there are countries that do not follow this trend.

Notably, countries like the United States, China, and Italy start with a relatively small Giant Component but experience rapid growth over time. On the other hand, some countries already have a Giant Component occupying more than 30% from the beginning.

These differences in the Giant Component trends reflect the varying dynamics of network formation and connectivity within each country's microbiome research landscape. Factors such as research collaborations, funding allocation, scientific infrastructure, and cultural aspects of the scientific community can contribute to these disparities. Understanding these variations provides insights into the development and structure of the microbiome research networks in different countries.

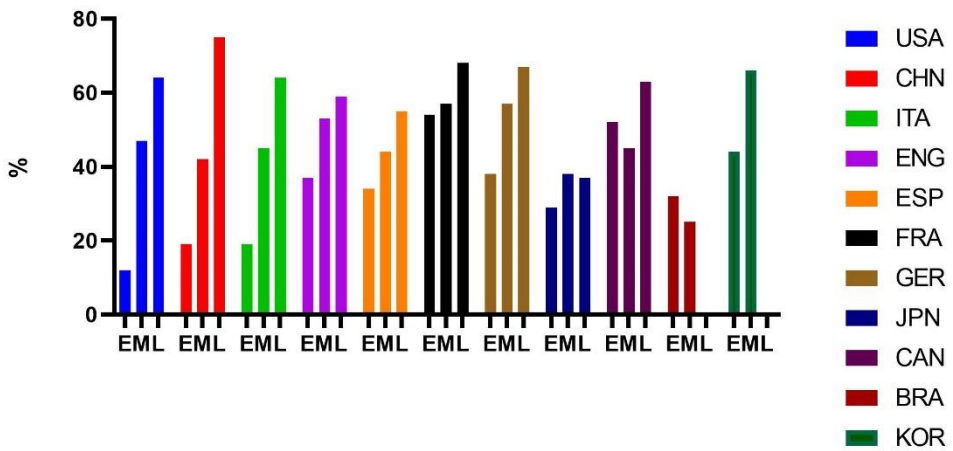

Figure III.4.1. Changes in Giant component of microbiome Co-Author Network at each stage

Y axis represents percentage of Giant component in total network.

E:Early phase, M:Mid phase, L:Late phase. See table 2

Identifying well-established networks in the early stage of the United States was challenging due to the prevalence of small clusters with only a few papers, resulting in a limited number of edges between clusters.. During this period, the researcher ('GORDON', 'JEFFREY I') emerged as a central figure in clusters with high centrality.

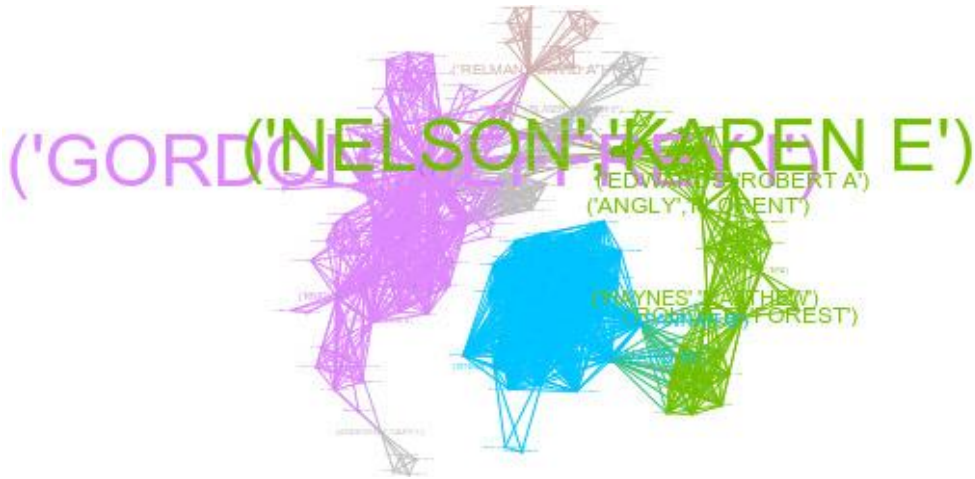

Figure III.4.2. The microbiome Co-Author Networks in United States Early Phase

Ratio of node and edges : 209 node (12.95%), 1708 edges (26.93%). Top Highest Betweeness Author : ('GORDON', 'JEFFREY I') 10987.96826, ('NELSON', 'KAREN E') 10644

From the mid-stage American networks, researchers in the field of "Analytical and mathematical methods in microbiome research" demonstrated high betweenness centrality. Notable researchers in this field include ('Knight, Rob').

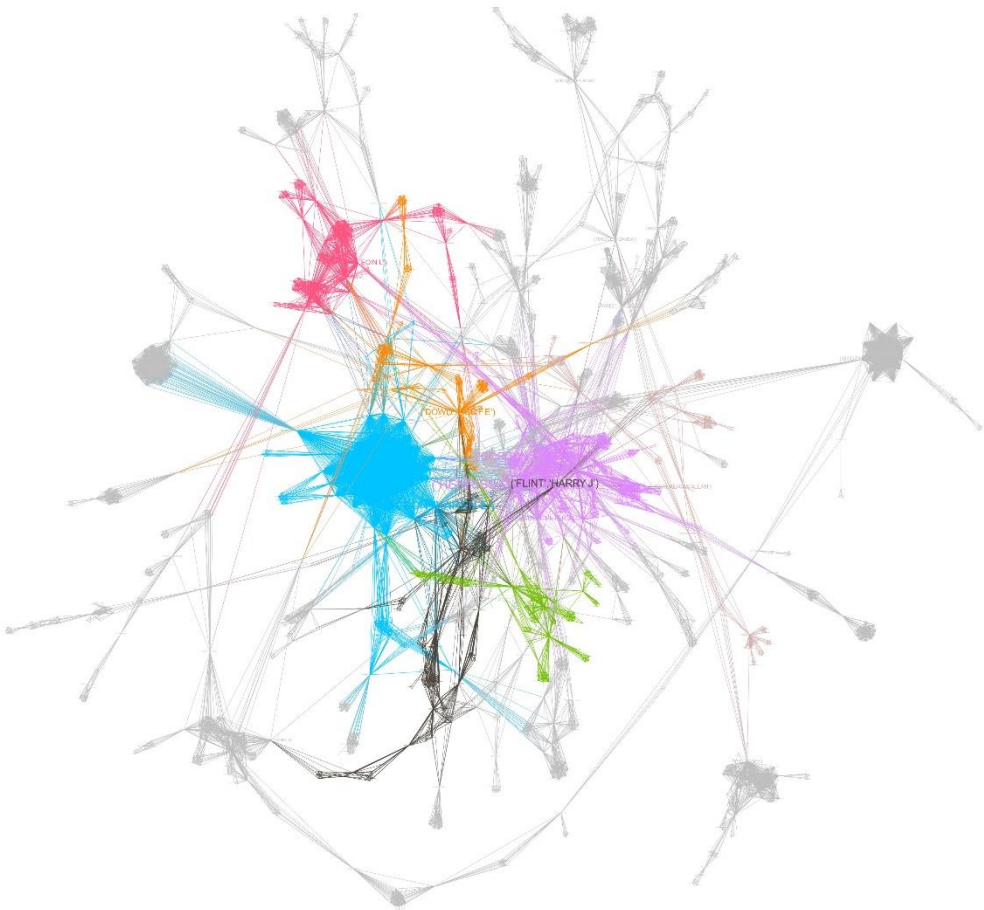

Figure III.4.3. The microbiome Co-Author Networks in United States Mid Phase

Ratio of node and edges : 2309 node (47.27%) ,18070 edges (52.21%). Top Highest Betweenness Author :('HENRISSAT', 'BERNARD') 417659.9625, ('NELSON', 'KAREN E') 385017.3332, ('GORDON', 'JEFFREY I') 360763.5466, ('FLINT', 'HARRY J') 341097.6293, ('KNIGHT','ROB') 335832.7091, ('DOWD', 'SCOT E') 313891.5051, ('BRODIE', 'EOIN L') 252384.1211, ('TURNBAUGH', 'PETER J') 239037.5876, ('WALKER', 'W ALLAN') 220230.7082

In the late stage, a fully developed large-scale network is clearly visible in the United States. Notable researchers during this period include ('Knight, Rob').

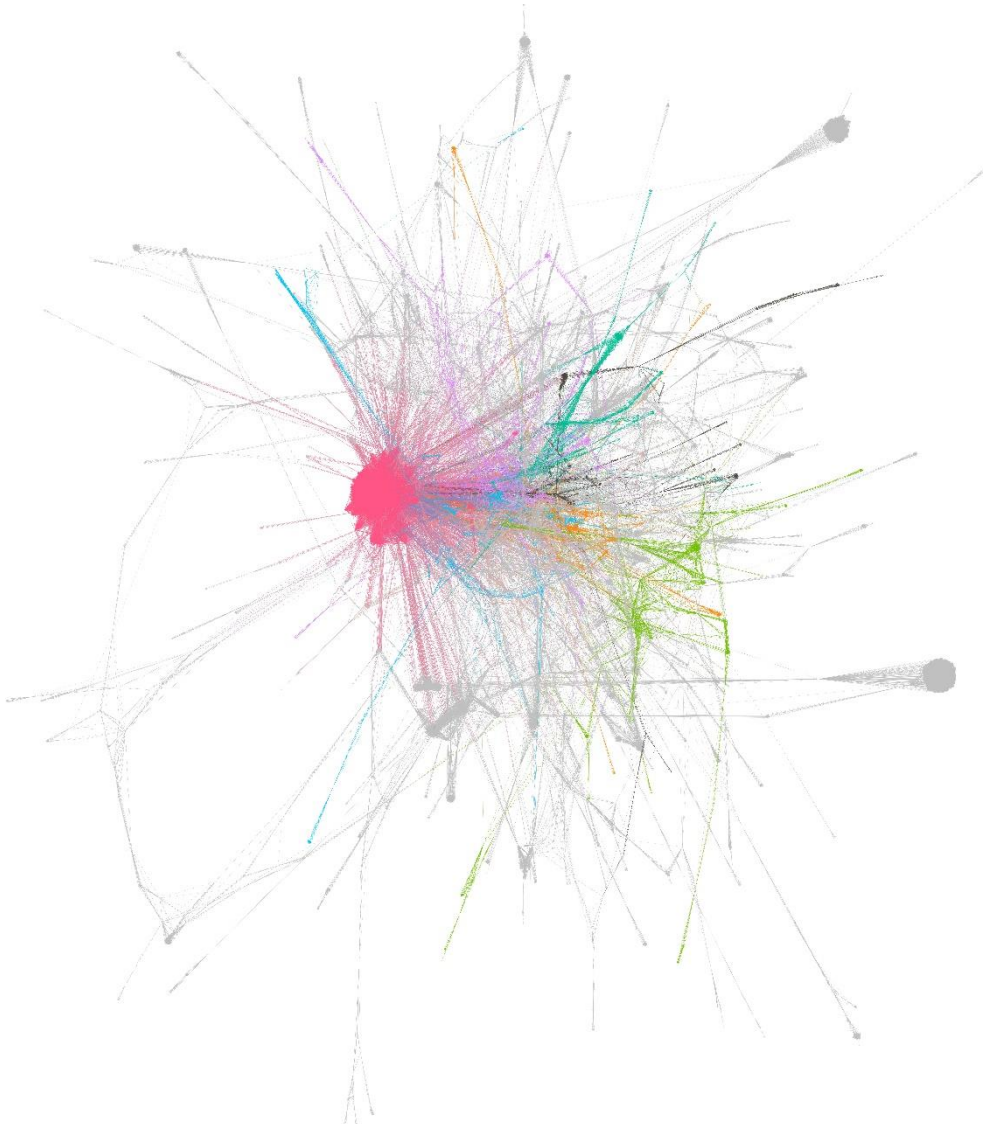

Figure III.4.4. The microbiome Co-Author Networks in United States Late Phase

Ratio of node and edges: 9329 node (64.09%) ,120544 edges (86.57%). Top Highest Betweenness Author: ('KNIGHT', 'ROB') 8790247.519

It is worth mentioning that all of the researchers mentioned above played significant

roles as authors in the 2012 Human Microbiome Project (HMP).(“Structure, function and diversity of the healthy human microbiome | Nature,” n.d.)

Based on these findings, it can be observed that the development of the network in the United States revolved around researchers in the field of "Analytical and mathematical methods in microbiome research."

Similar to the Unites States, it was also challenging to identify significant networks in China in the early stage.

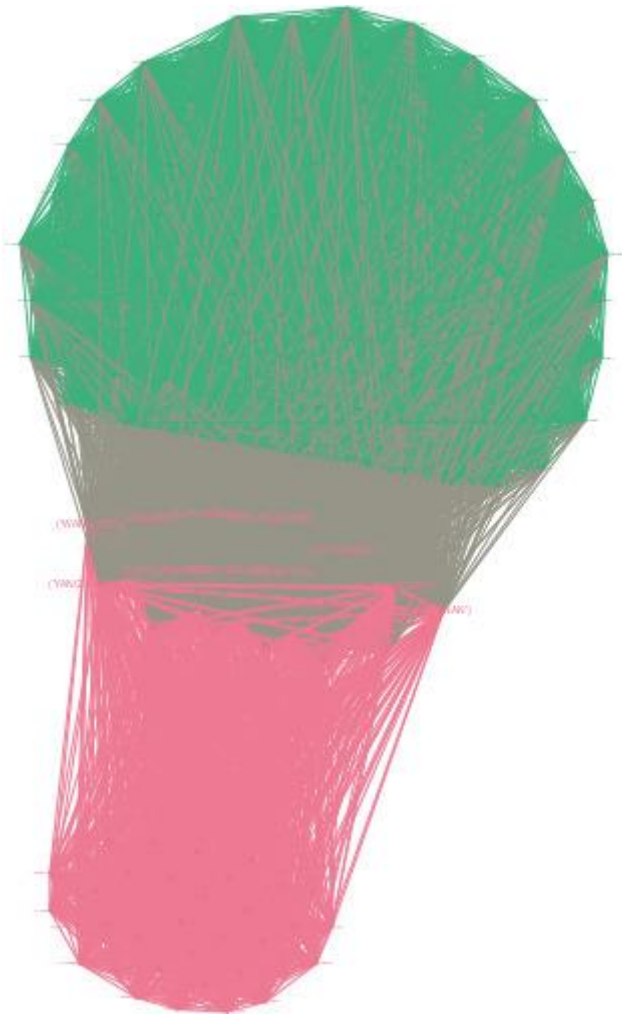

Figure III.4.5. The microbiome Co-Author Networks in China Early Phase

Ratio of node and edges: 159 node (19.9%) ,8745 edges (74.9%). Top Highest Betweenness Author : ('LI', 'RUIQIANG'), ('ZHU', 'HONGMEI'), ('JIAN', 'MIN'), ('ZHOU','YAN'),('CAO','JIANJUN'),('WANG','BO'),('YU','CHANG'),('LIANG','HUIQING'),('ZHENG','HUISONG'),('LI','YINGRUI'),('QIN','NAN'),('KRISTIANSEN',KARSTEN'),('ZHANG','XIUQING'),('LI','SONGGANG'),('YANG','HUANMING'), ('WANG', 'JIAN'), ('WANG','JUN') betweenness is equal as 224.4705882

In the mid stage, like the United States, the formation of networks in China can be clearly observed. Prominent doctors such as Li Lanjuan, who developed the Li-NBAL artificial liver support system used to sustain the lives of people suffering from acute liver failure, appear in large clusters.

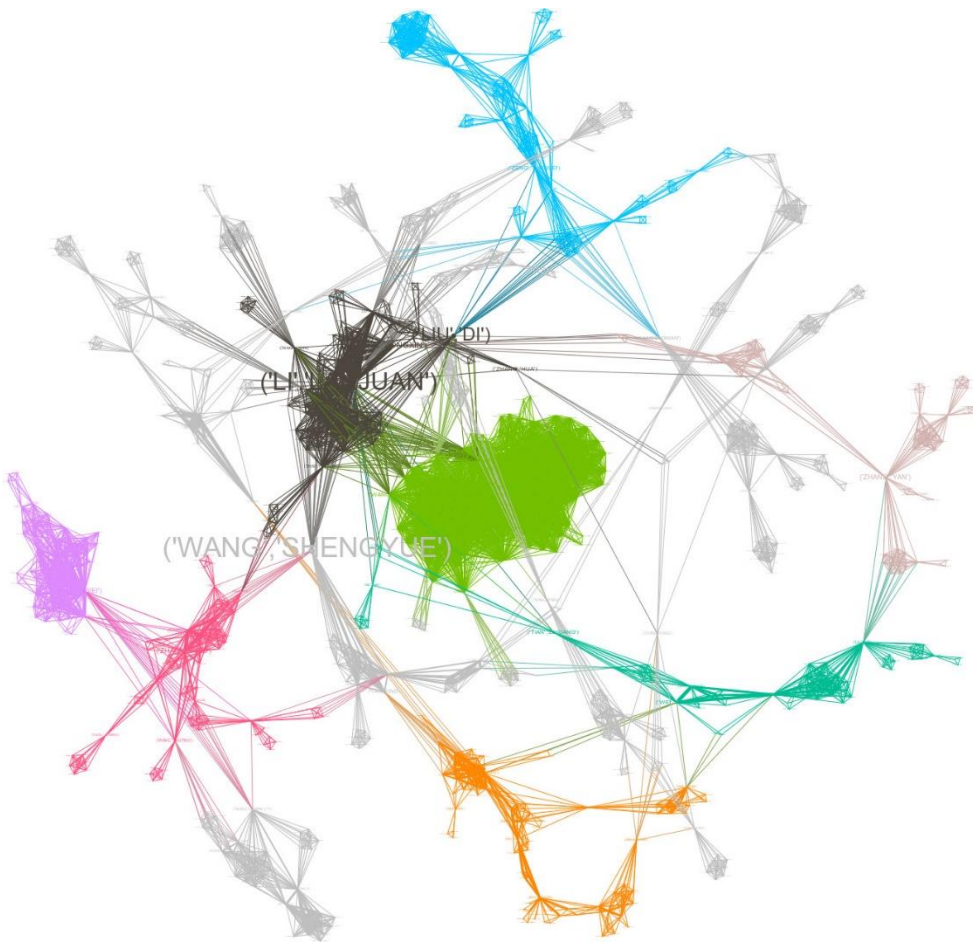

Figure III.4.6. The microbiome Co-Author Networks in China Mid Phase

Ratio of node and edges: 1090 node (42.07%) ,8719 edges (60.68%). Top Highest Betweenness Author :('WANG', 'SHENGYUE') 183986.7062 ('LI', 'LANJUAN') 169767.7221 ('LIU', 'DI') 136447.5248

In the late stage, the size of the network and the presence of researchers in the field of "Analytical and mathematical methods in microbiome research" are consistently maintained, mirroring the findings in the United States.

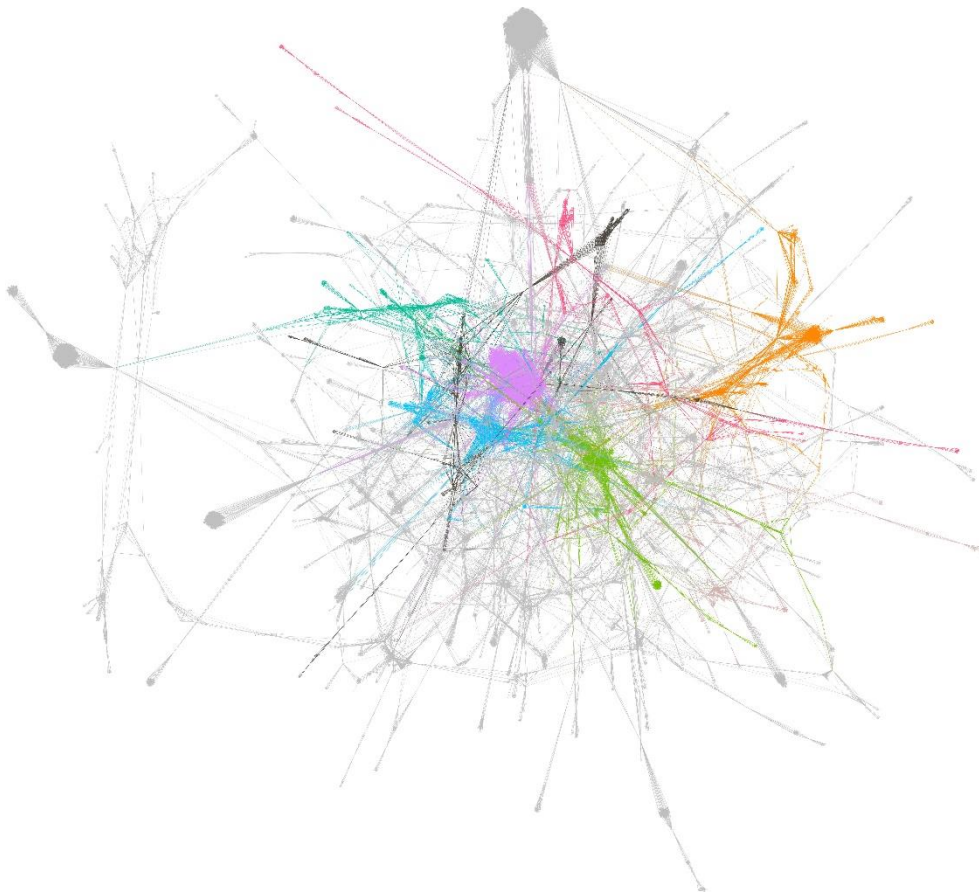

Figure III.4.7. The microbiome Co-Author Networks in China Late Phase

Ratio of node and edges: 6301 node (75.6%), 49882 edges (75.34%). Top Highest Betweenness Author: ('LI', 'JUN') 2338798.275('YIN','YULONG')2101550.311 ('BLACHIER', 'F') 1890944 ('WU', 'X') 1771513.411 ('WANG', 'JUN') 1621138.612 ('LI', 'YAN') 1234311.576

Figure III.4.8. The microbiome Co-Author Networks in Canada Early Phase

Ratio of node and edges: 1029 node (52.23%) ,38950 edges (91.59%). Top Highest Betweenness Author : ('ALLENVERCOE', 'EMMA') 272516.1195 ('FLINT', 'HARRY J') 247097.738

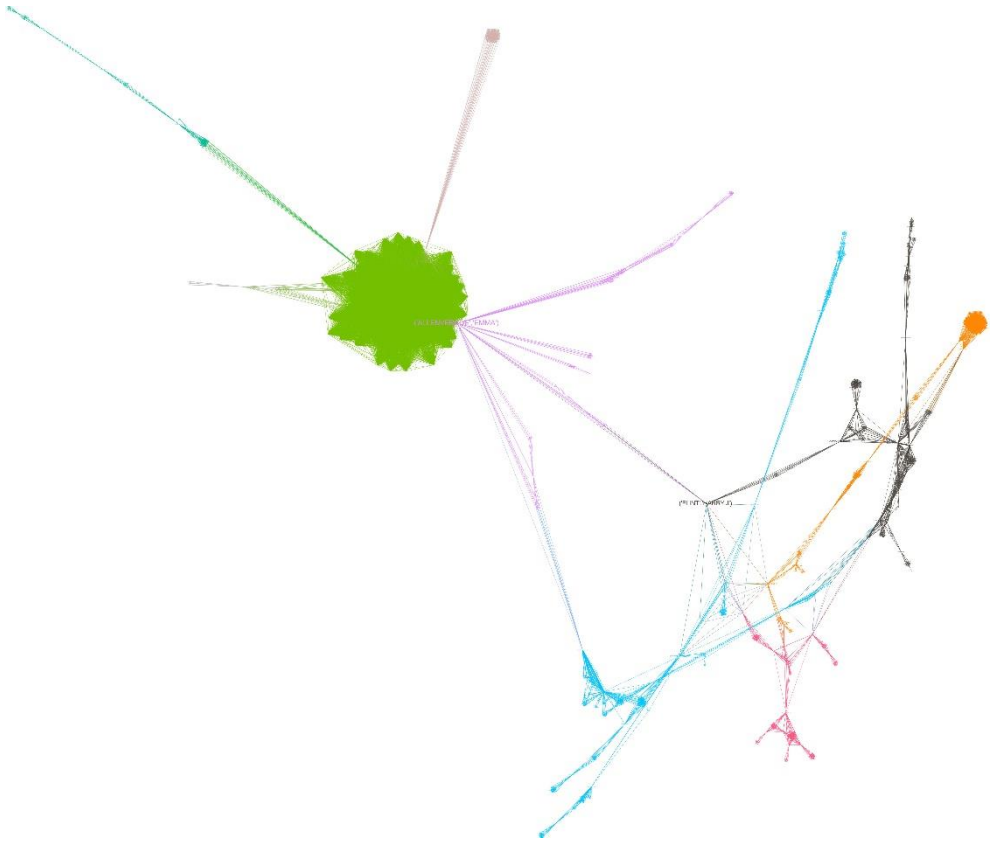

Figure III.4.9. The microbiome Co-Author Networks in Canada Mid Phase

Ratio of node and edges: 2187 node (45.52%) ,21195 edges (49.28%). Top Highest Betweenness Author : ('KNIGHT', 'ROB') 422698.4923 ('GUTTMAN', 'DAVID S') 330088.3185 ('CHEN', 'RUI') 328490.911 ('FINLAY', 'B BRETT') 293616.5816 ('ABBOTT', 'D WADE') 255801.841 ('SENSEN', 'CHRISTOPH W') 236208 ('GRUNINGER', 'ROBERT J') 217151.6656 ('GRUBE', 'MARTIN') 216528 ('REID', 'GREGOR') 213577.2238

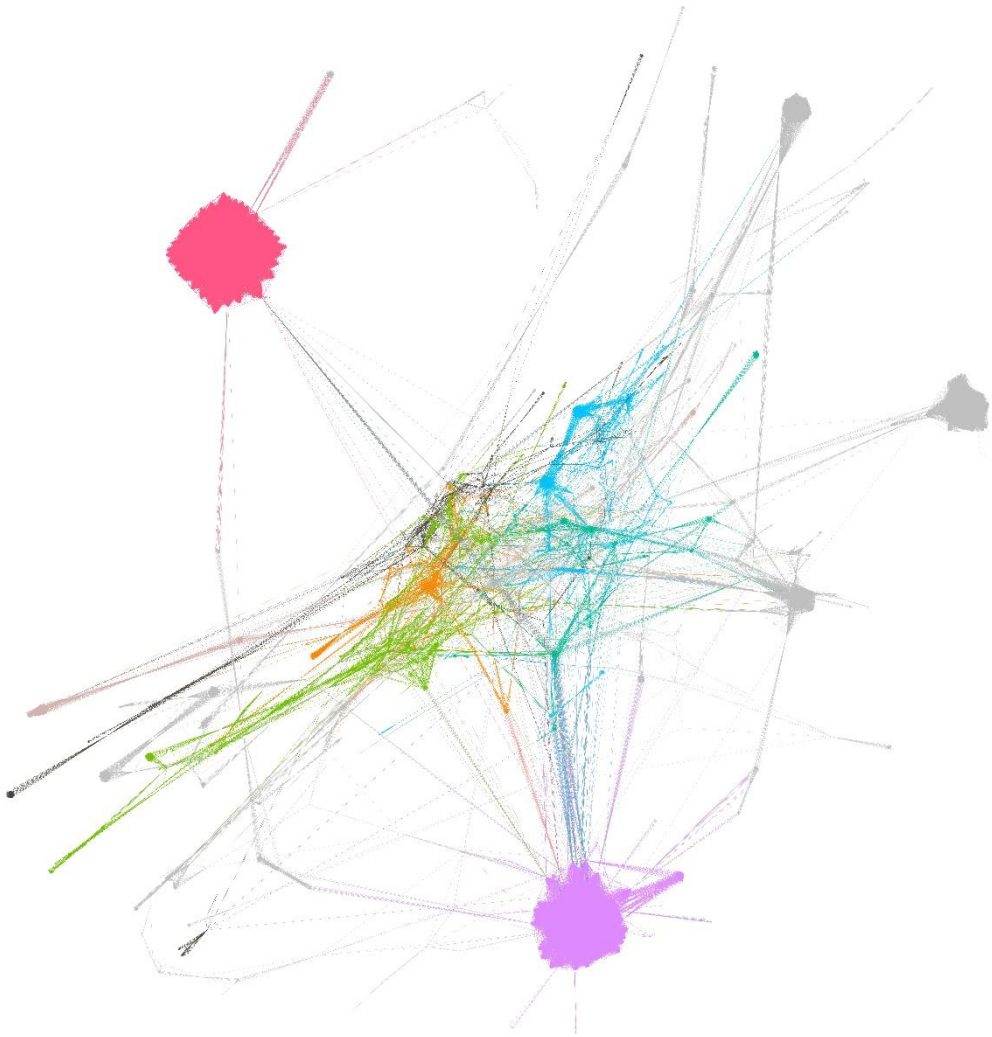

Figure III.4.10. The microbiome Co-Author Networks in Canada Late Phase

Ratio of node and edges: 6132 node (63.76%) ,201713 edges (88.49%). Top Highest Betweenness Author : ('JIA', 'WEI') 2693632.627 ('SURETTE', 'MICHAEL G') 2256873.032 ('XAVIER', 'RAMNIK J') 2024373.817 ('ALLENVERCOE', 'EMMA') 1856371.913 ('KNIGHT', 'ROB') 1512513.714 ('WALTER', 'JENS') 1377945.274

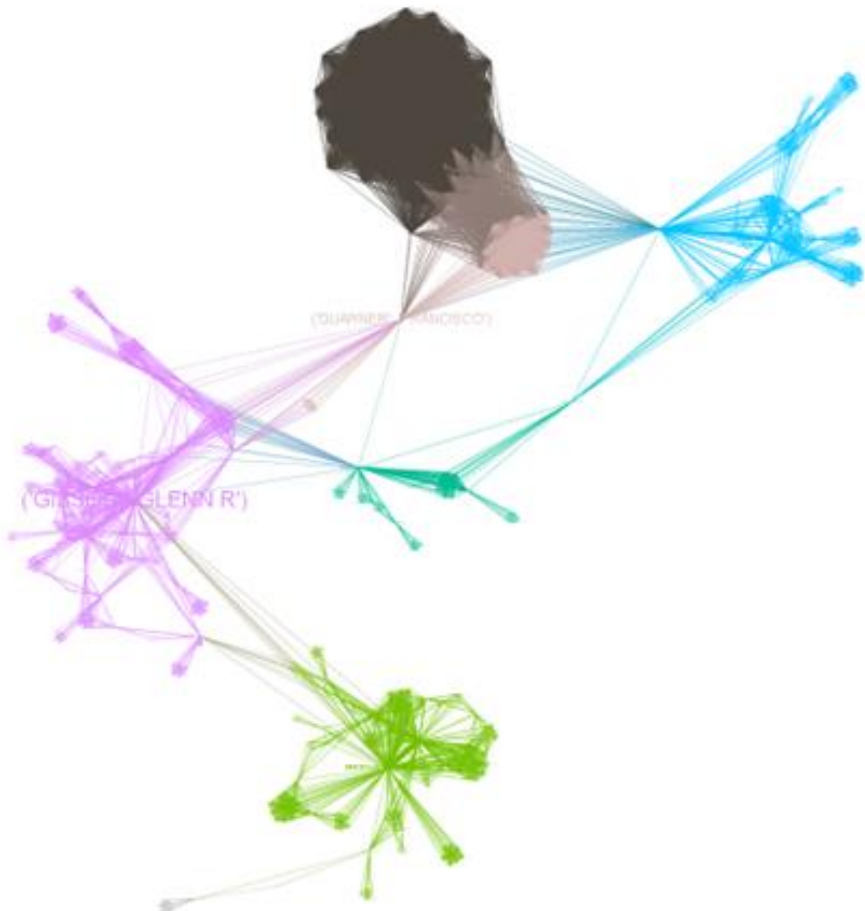

Figure III.4.11. The microbiome Co-Author Networks in England Early Phase

Ratio of node and edges: 700 node (37.76%) ,12090 edges (66.39%). Top Highest Betweenness Author : ('GIBSON', 'GLENN R') 138551.1576 ('GUARNER', 'FRANCISCO') 105420.4848

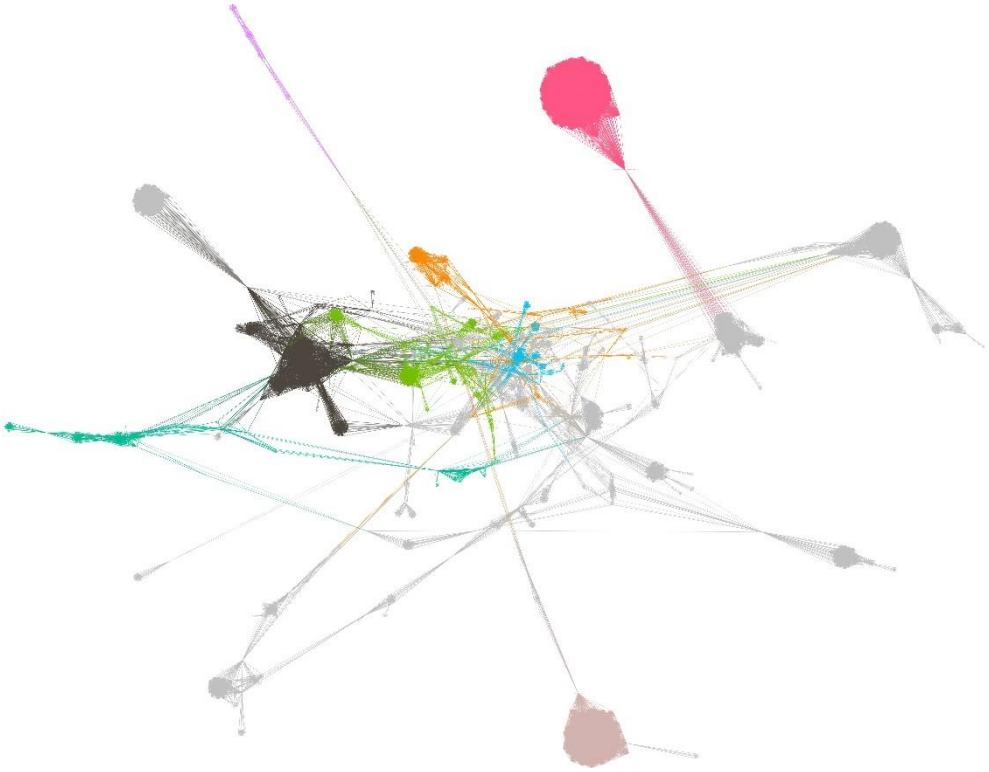

Figure III.4.12. The microbiome Co-Author Networks in England Mid Phase

Ratio of node and edges: 2498 node (53.59%), 42192 edges (73.74%). Top Highest Betweenness Author: ('FLINT', 'HARRY J') 1070629.18

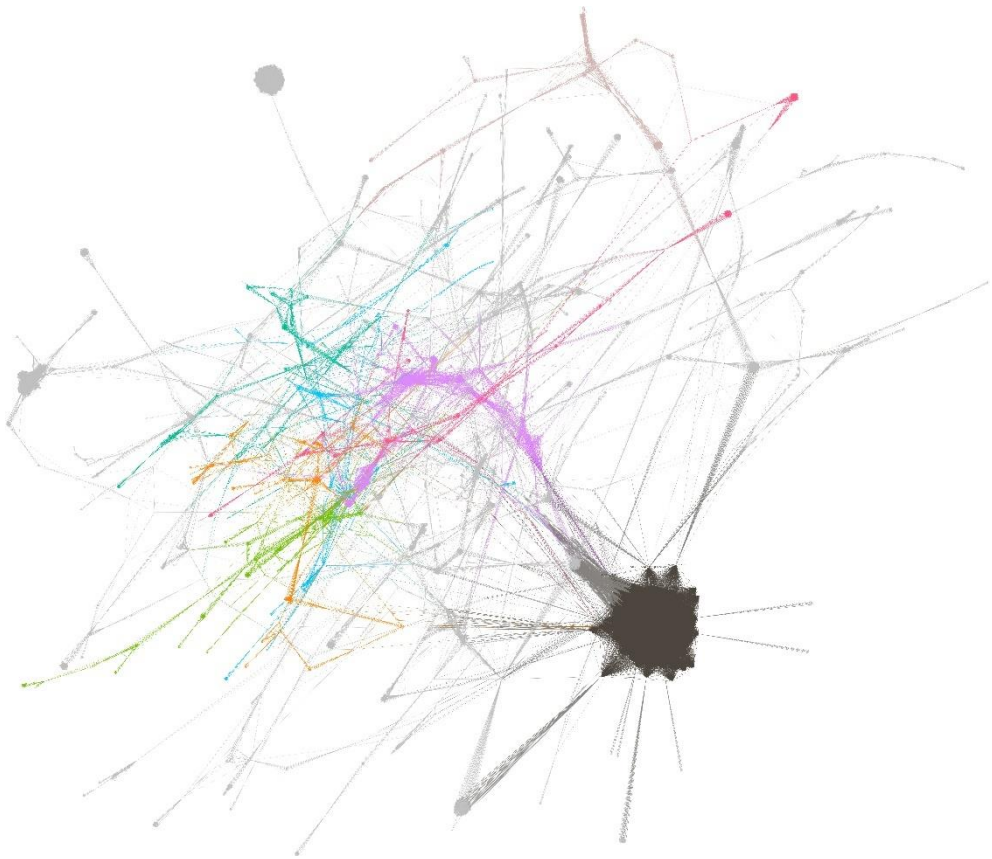

Figure III.4.13. The microbiome Co-Author Networks in England Late Phase

Ratio of node and edges: 5789 node (59.11%) ,102072 edges (80.11%). Top Highest Betweenness Author :('MARCHESI', 'JULIAN R') 1761056.026 ('MCDONALD', 'JAMES E') 1700174.796 ('KNIGHT', 'ROB')1356902.349 ('HOLMES', 'ELAINE') 1300277.81 ('NOLAN', 'MATTHEW J') 1144765.342 ('WANG', 'JUN') 1073760.374 ('WALKER', 'ALAN W') 994271.8265 ('GROSSART', 'HANSPETER') 986313.3233 ('BOKULICH', 'NICHOLAS A') 962573.0617 ('FLINT', 'HARRY J') 929871.1966 ('WADE', 'WILLIAM G') 912984.2274 ('JACKSON', 'MATTHEW A') 903968.2947 ('PARKHILL', 'JULIAN') 900800.1148 ('SCHLOTER', 'MICHAEL') 884646.9672

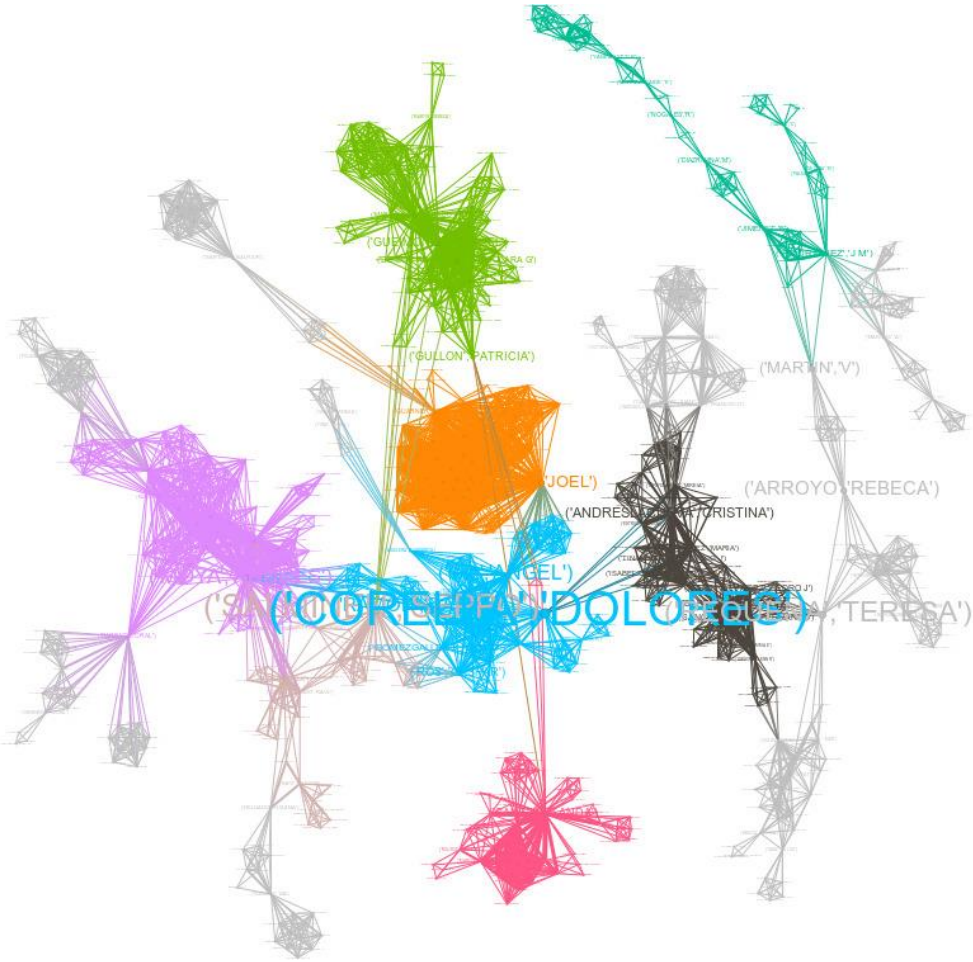

Figure III.4.14. The Co-Author Networks in Spain Early Phase

Ratio of node and edges: 747 node (34.89%) ,4890 edges (42.73%). Top Highest Betweenness Author: ('CORELLA', 'DOLORES') 129428 ('SALMINEN', 'SEPPO') 98900.39197 ('REQUENA', 'TERESA') 89617.84137 ('MOYA', 'ANDRES') 70721.05694 ('GIL', 'ANGEL') 67228.60952

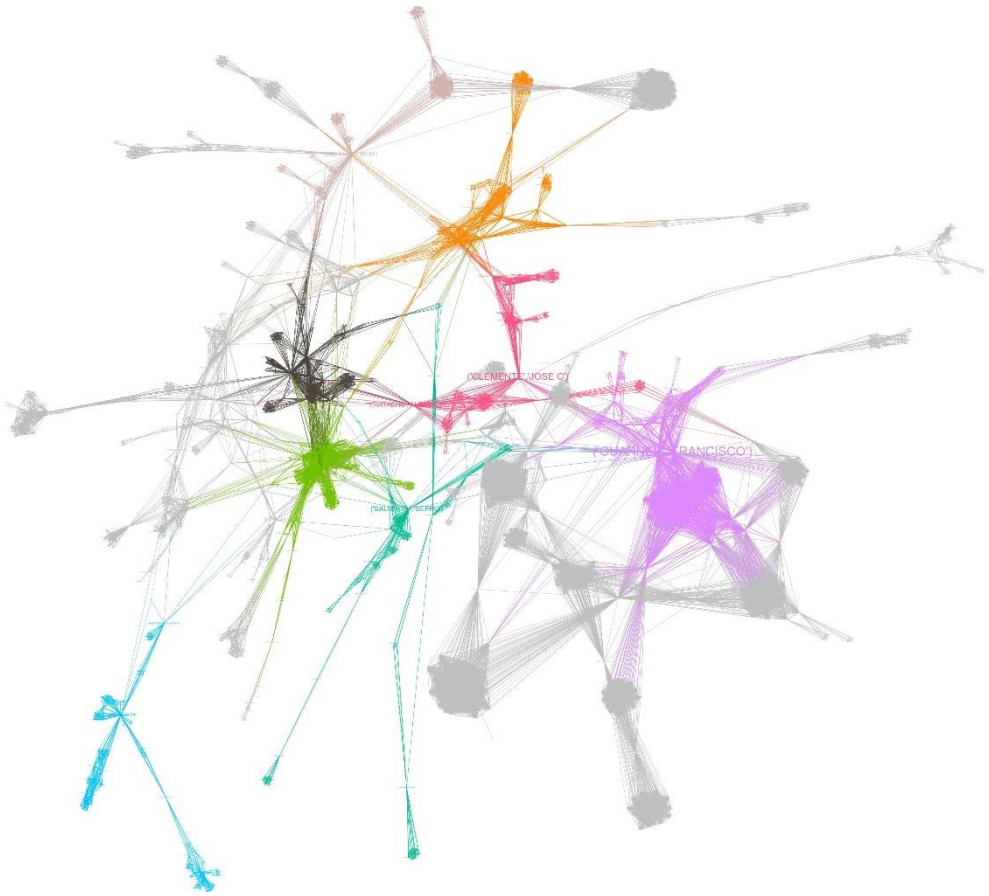

Figure III.4.15. The microbiome Co-Author Networks in Spain Mid Phase

Ratio of node and edges: 1953 node (44.81%), 20064 edges (47.94%). Top Highest Betweenness Author :('CARMEN COLLADO', 'MARIA') 1530416.185 ('MOYA', 'ANDRES') 1233055.533 ('MIRA', 'ALEX') 993292.9305 ('GUARNER', 'FRANCISCO') 540317.0119 ('CLEMENTE', 'JOSE C') 430019.7564 ('SALMINEN','SEPPO') 356168.3186 ('ARTACHO','ALEJANDRO') 337544.0809 ('DEL CAMPO', 'ROSA') 274117.946

Figure III.4.16. The microbiome Co-Author Networks in Spain Late Phase

Ratio of node and edges: 4673 node (55.91%), 99420 edges (81.47%). Top Highest Betweenness Author: ('CARMEN COLLADO', 'MARIA') 1530416.185 ('MOYA', 'ANDRES') 1233055.533 ('MIRA', 'ALEX') 993292.9305

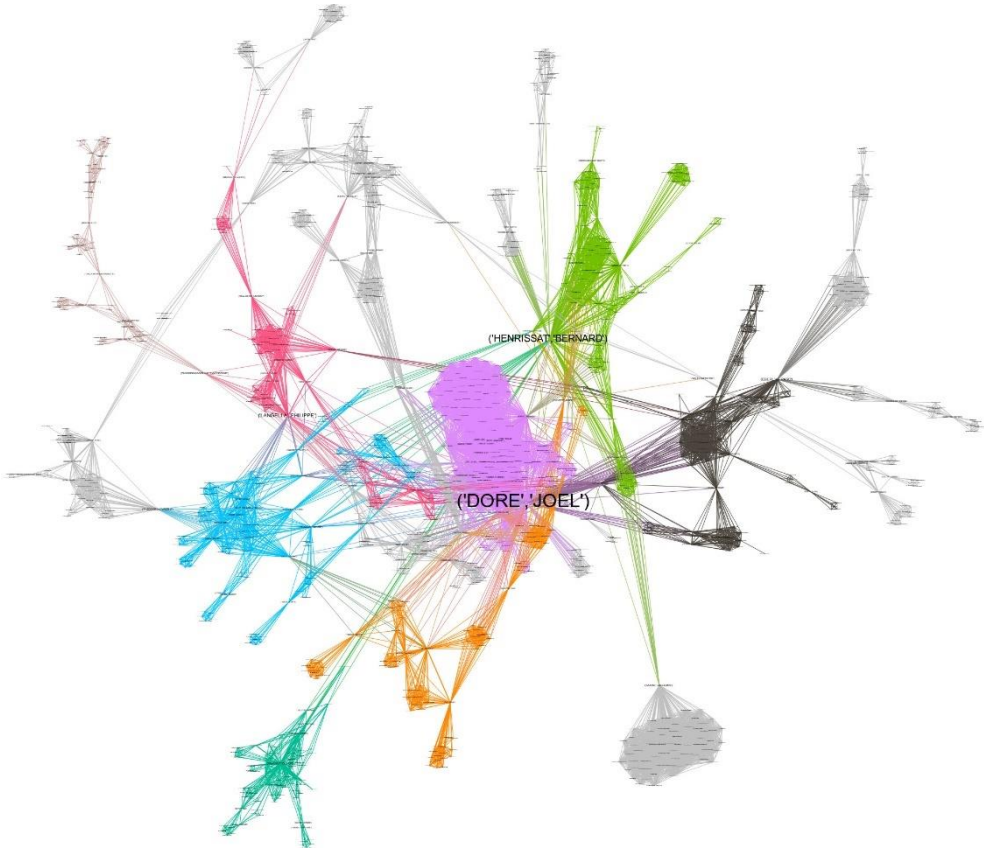

Figure III.4.17. The microbiome Co-author networks in France Early Phase

Ratio of node and edges: 1305 node (54.4%) ,12661 edges (73.98%). Top Highest Betweenness Author: ('DORE', 'JOEL') 418192.4267 ('HENRISSAT', 'BERNARD') 262278.4559

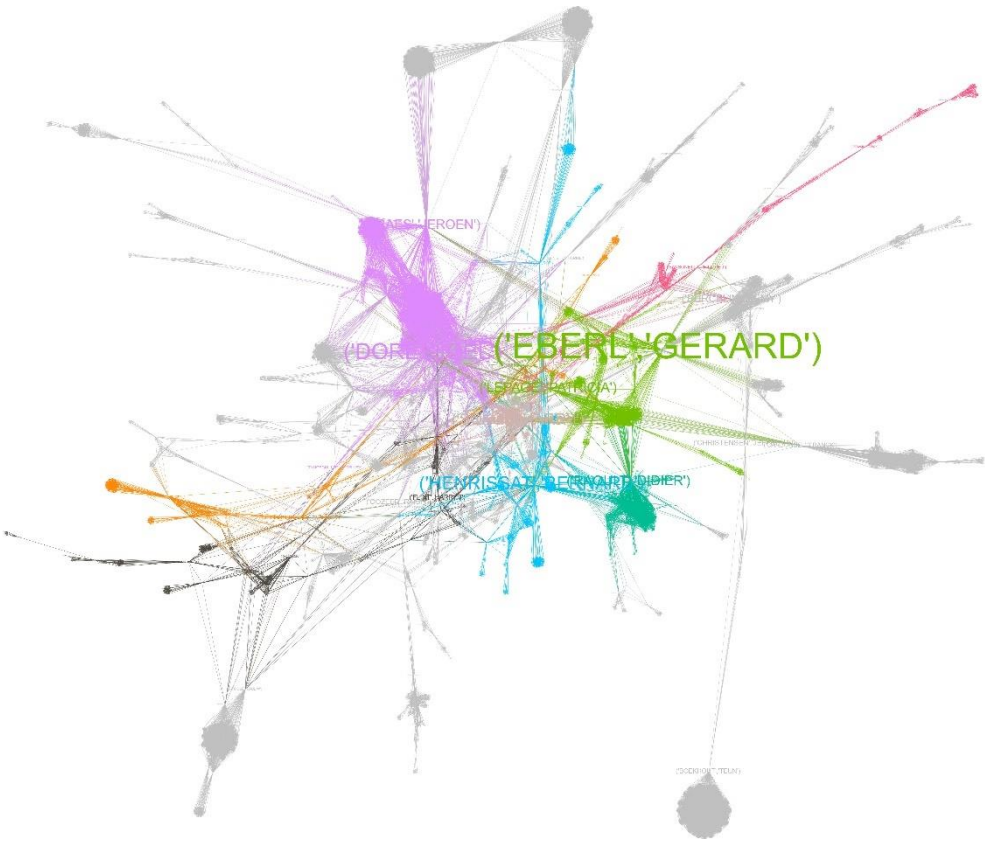

Figure III.4.18. The microbiome Co-author networks in France Mid Phase

Ratio of node and edges: 3320 node (57.25%) ,40624 edges (68.94%). Top Highest Betweenness Author: ('EBERL', 'GERARD') 967826.9721 ('DORE', 'JOEL') 735350.9195 ('HENRISSAT', 'BERNARD') 648278.1184 ('SOKOL', 'HARRY') 578626.3509 ('RAOULT', 'DIDIER') 536945.0636 ('LEPAGE', 'PATRICIA') 535598.5065 ('RAES', 'JEROEN') 521820.3113

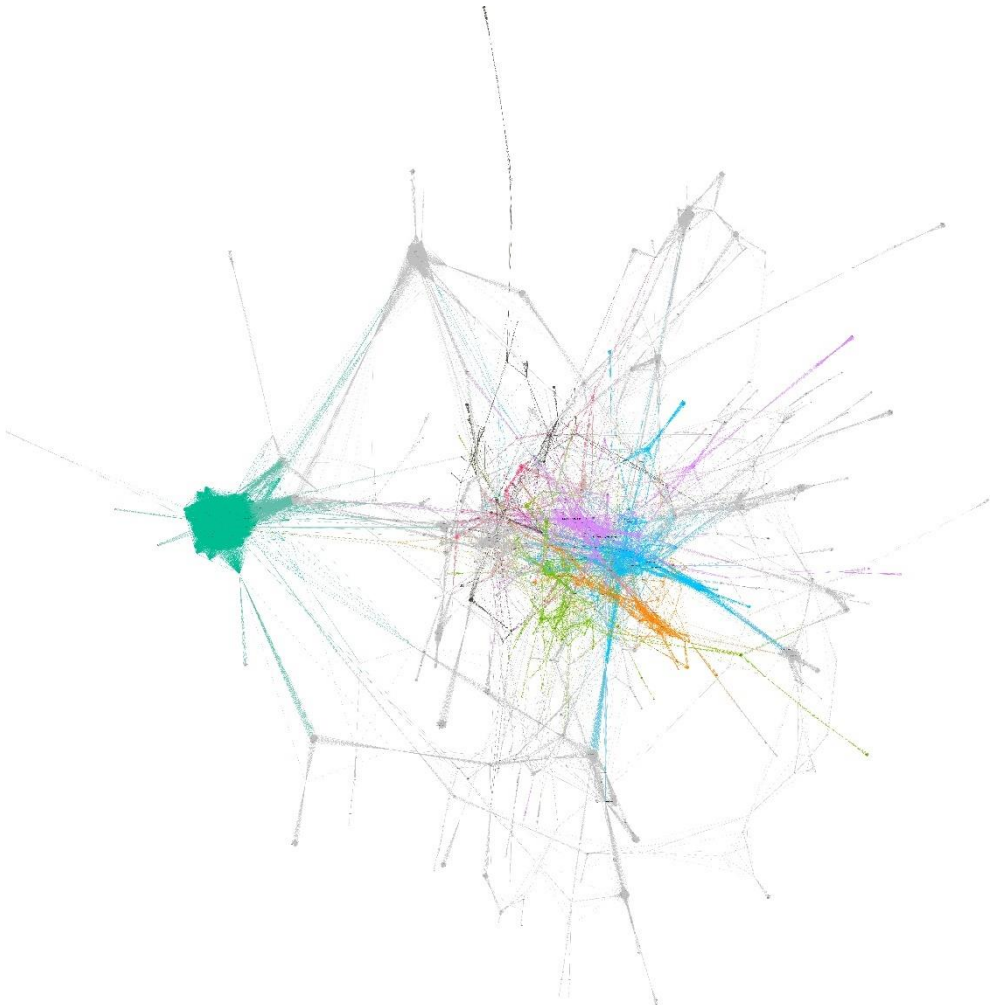

Figure III.4.19. The microbiome Co-author networks in France Late Phase

Ratio of node and edges: 7666 node (68.31%) ,127247 edges (79.89%). Top Highest Betweenness Author :('SOKOL', 'HARRY') 3574401.449 ('LANGELLA', 'PHILIPPE') 3327407.342 ('HENRISSAT', 'BERNARD') 2268113.031 ('PONS', 'NICOLAS') 1837730.506 ('RAOULT', 'DIDIER') 1837111.611

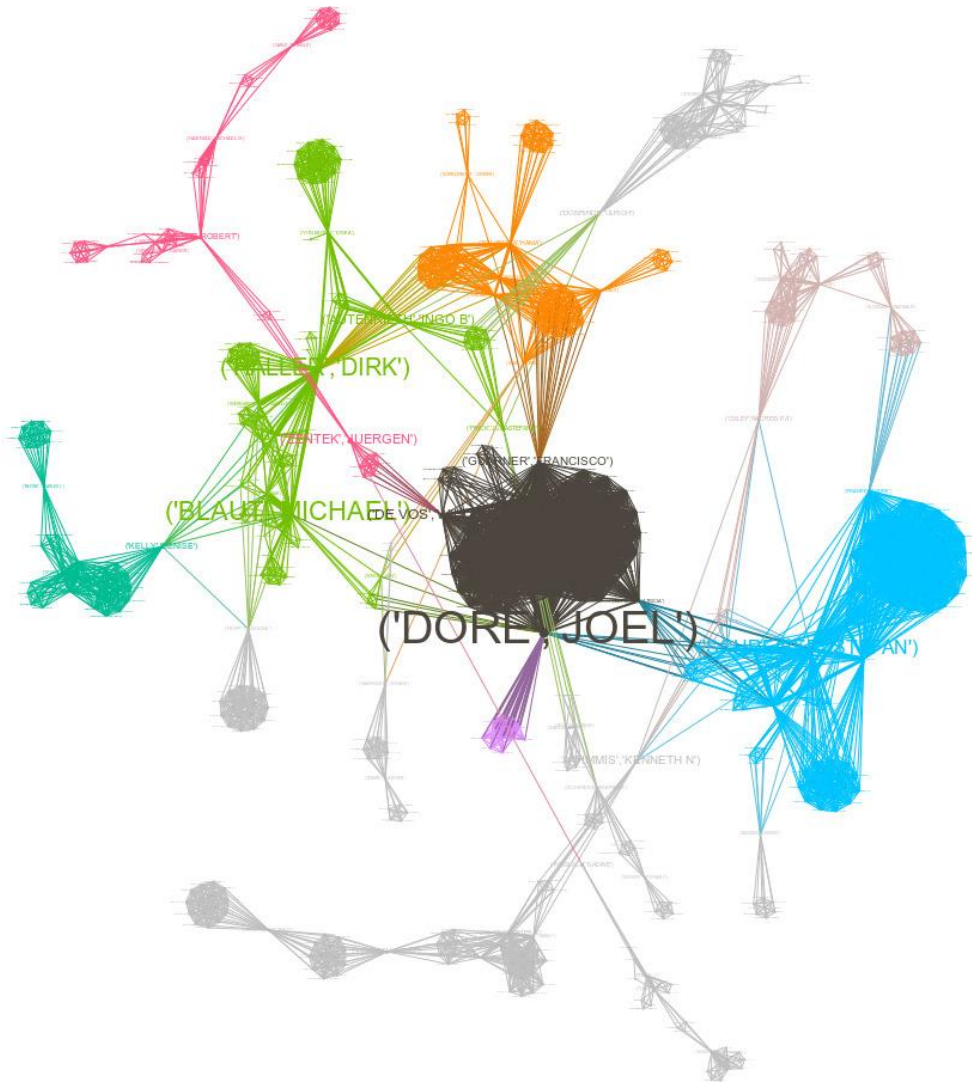

Figure III.4.20. The microbiome Co-Author Networks in Germany Early Phase

Ratio of node and edges: 645 node (38.23%) ,5871 edges (56.52%). Top Highest Betweenness Author: ('DORE', 'JOEL') 101456.4526 ('BLAUT', 'MICHAEL') 68245.2894 ('HALLER', 'DIRK') 64749.60325 ('SCHREIBER', 'STEFAN') 53865.99001

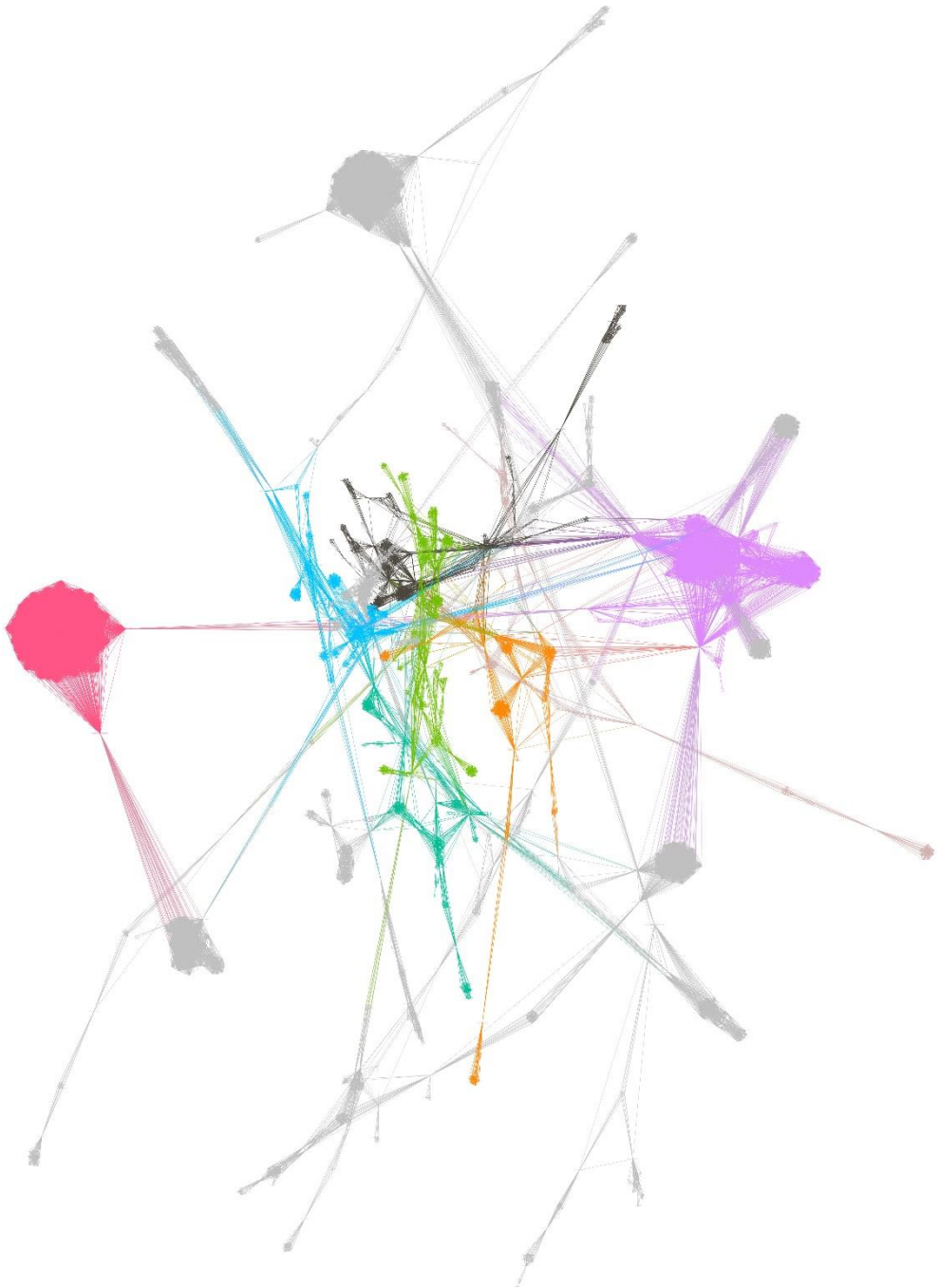

Figure III.4.21. The microbiome Co-Author Networks in Germany Mid Phase

Ratio of node and edges: 2593 node (57.27%) ,38375 edges (82.28%). Top Highest Betweenness Author: ('WANG', 'JUN') 913125.7845 ('BLAUT', 'MICHAEL') 527332.6383 ('STECHER', 'BAERBEL') 502795.6805

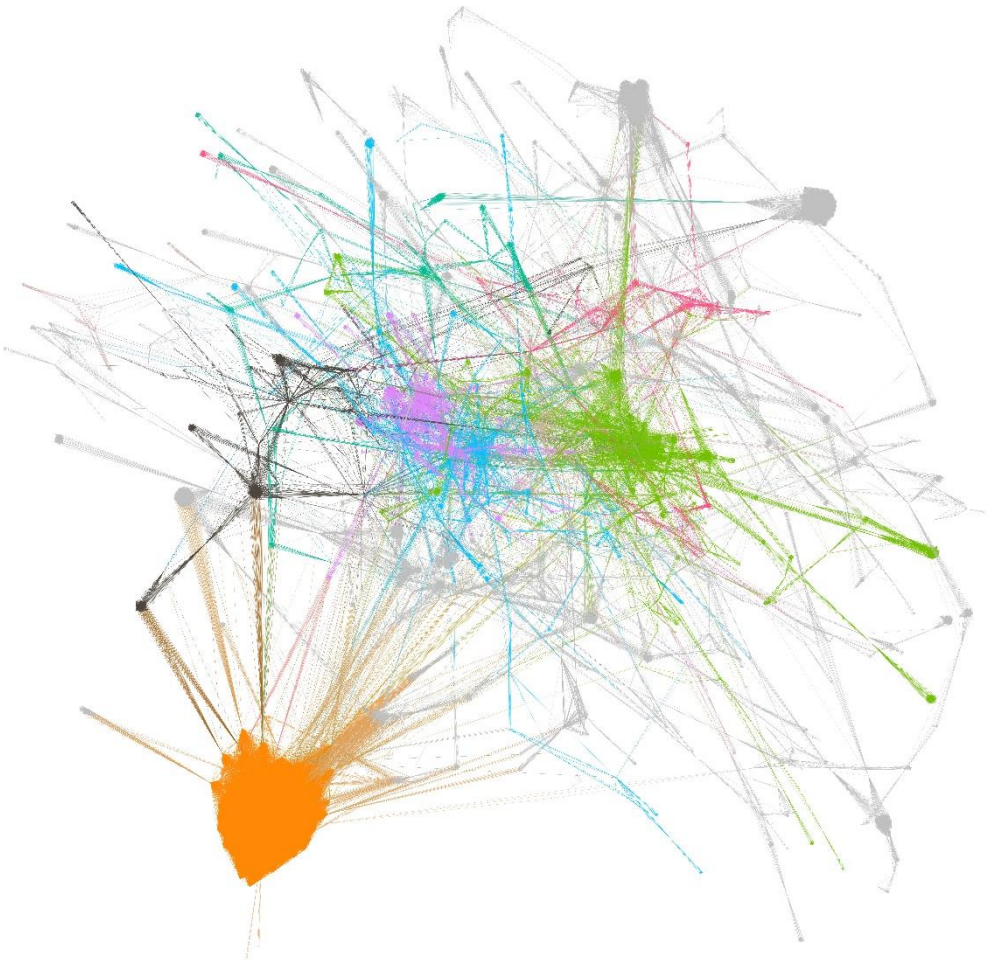

Figure III.4.22. The microbiome Co-Author Networks in Germany Late Phase

Ratio of node and edges: 7091 node (67.38%) ,117949 edges (85.38%). Top Highest Betweenness Author :('BAINES', 'JOHN F') 3255164.326 ('WANG', 'JUN') 1704137.25

Figure III.4.23. The microbiome Co-Author Networks in Italia Early Phase

Ratio of node and edges: 355 node (19.44%) ,2170 edges (21.64%). Top Highest Betweenness Author: ('BRIGIDI', 'PATRIZIA') 24267.0852 ('MATTEUZZI', 'DIEGO') 18785 ('DI GIOIA', 'DIANA') 16447 ('NDAGIJIMANA', 'MAURICE') 15054.52069 ('OTOOLE', 'PAUL W') 13738.06741 ('MOGNA', 'LUCA') 12935.33333

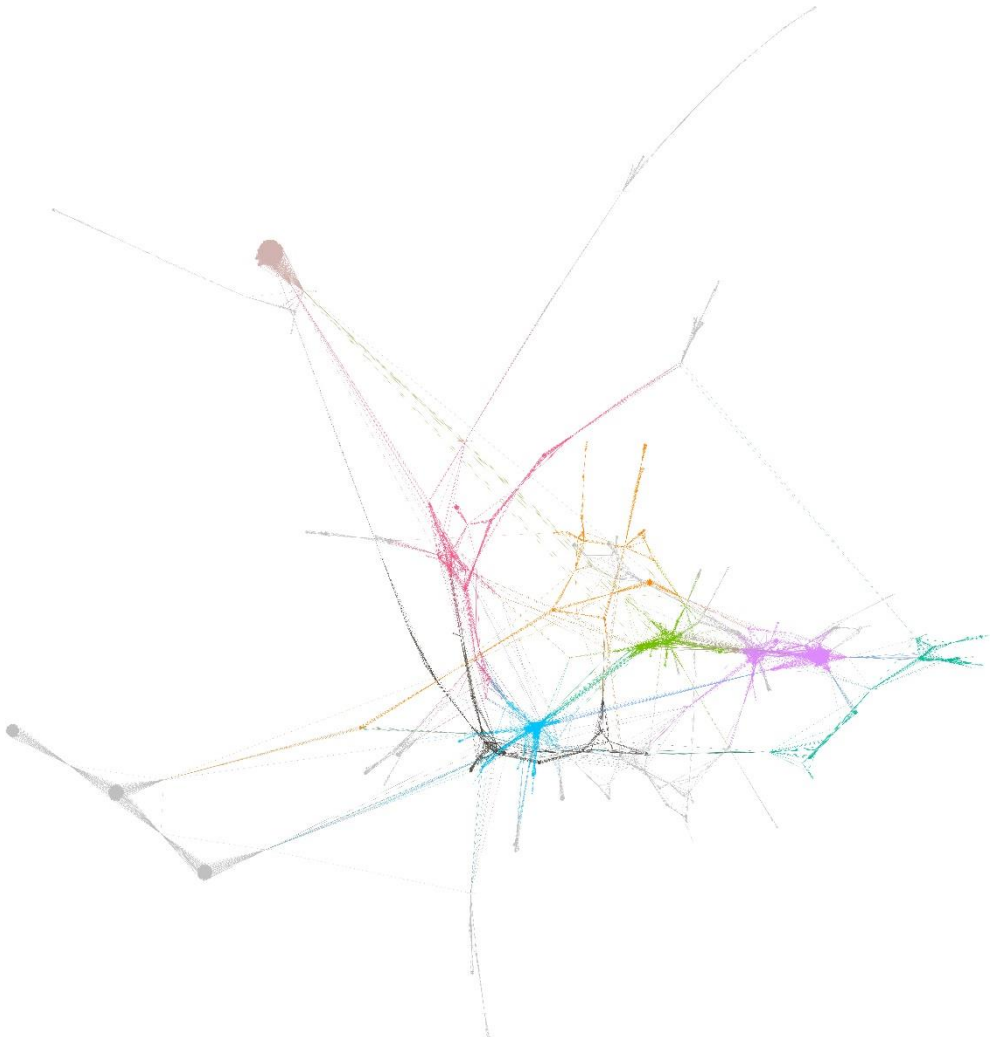

Figure III.4.24. The microbiome Co-Author Networks in Italia Mid Phase

Ratio of node and edges: 2317 node (45.86%) ,25979 edges (50.74%). Top Highest Betweenness Author : ('BRIGIDI', 'PATRIZIA') 388801.8526 ('FOSSO', 'BRUNO') 333380.546 ('BARBARA', 'GIOVANNI') 282416.8205 ('CARDINALI', 'GIANLUIGI') 269028.643 ('VENTURA', 'MARCO') 245046.6407 ('DE VOS', 'WILLEM') 228336.7673 ('GILBERT', 'JACK A') 210220.9641 ('ERCOLINI', 'DANILO') 208111.9077 ('SEGATA', 'NICOLA') 204633.0615 ('DI CAGNO', 'RAFFAELLA') 204264.8547 ('CANANI', 'ROBERTO BERNI') 198419.2825 ('NIESLER', 'BEATE') 198402.9214 ('GRIECO', 'FRANCESCO') 196064.25

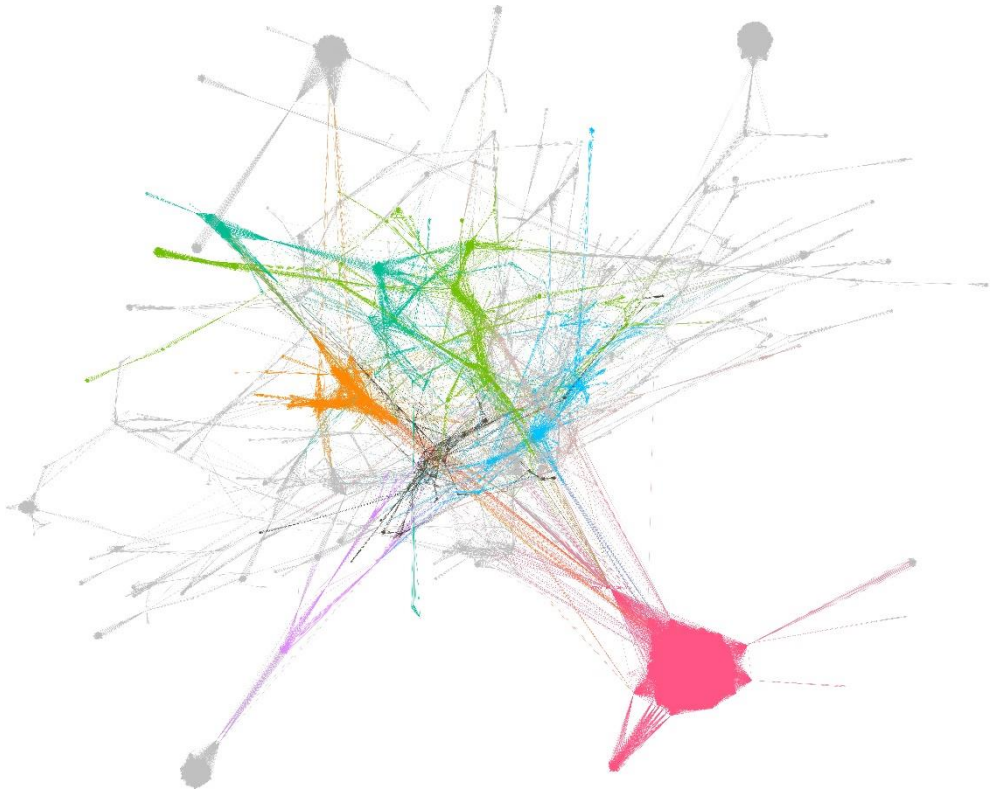

Figure III.4.25. The microbiome Co-Author Networks in Italia Late Phase

Ratio of node and edges: 6676 node (64.08%) ,117339 edges (83.23%). Top Highest Betweenness Author : ('ERCOLINI', 'DANILO') 2776445.573 ('SEGATA', 'NICOLA') 2071940.886 ('PUTIGNANI', 'LORENZA') 1938509.702 ('BRIGIDI', 'PATRIZIA') 1407609.092

Figure III.4.26. The microbiome Co-Author Networks in Japan Early Phase

Ratio of node and edges: 209 node (29.6%) ,1147 edges (37.83%). Top Highest Betweenness Author: ('ITOH', 'KIKUJI') 11617.5 ('BENNO', 'YOSHIMI') 9814.875 ('KUWAHARA', 'TOMOMI') 9660

Figure III.4.27. The microbiome Co-Author Networks in Japan Mid Phase

Ratio of node and edges: 452 node (38.4%) , 3277 edges (53.38%). Top Highest Betweenness Author: ('TAKEDA', 'KIYOSHI') 51009.88889 ('ITOH', 'KIKUJI') 49859.5 ('YAMAMOTO', 'MASAHIRO') 40780.53968 ('MATSUMOTO', 'MITSUHARU') 33648 ('HATTORI', 'MASAHIRA') 27594

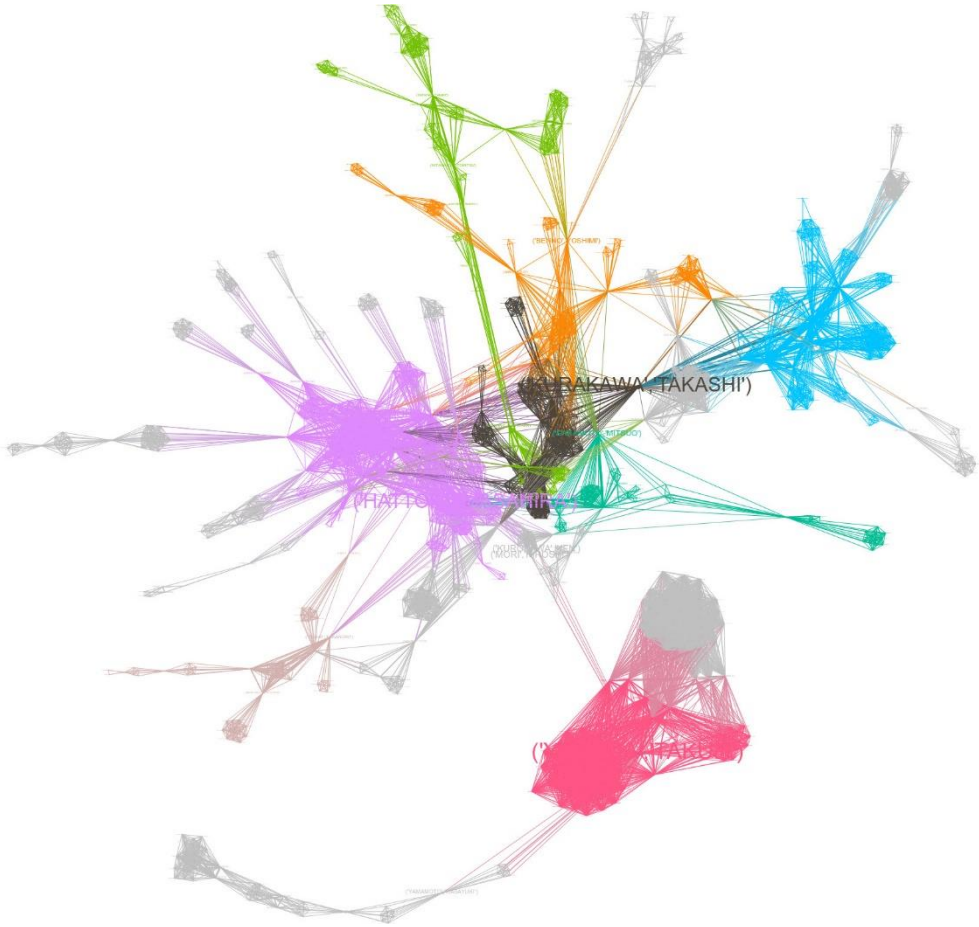

Figure III.4.28. The microbiome Co-Author Networks in Japan Late Phase

Ratio of node and edges: 1188 node (37.94%) ,10880 edges (31.19%). Top Highest Betweenness Author : ('YAMADA', 'TAKUJI') 183168.5357 ('KURAKAWA', 'TAKASHI') 166980.4354 ('HATTORI', 'MASAHIRA') 166924.3761 ('OSHIMA', 'KENSHIRO') 110866.1343 ('HONDA', 'KENYA') 107618.7476

The overall network changes in South Korea show little difference compared to other countries in terms of the timing of stages. However, in the late stage, the overall size of the network is approximately half that of other countries. The exposure of researchers in the field of "Analytical and mathematical methods in microbiome research" in the top authors of the modules appears to be relatively smaller compared to other countries.

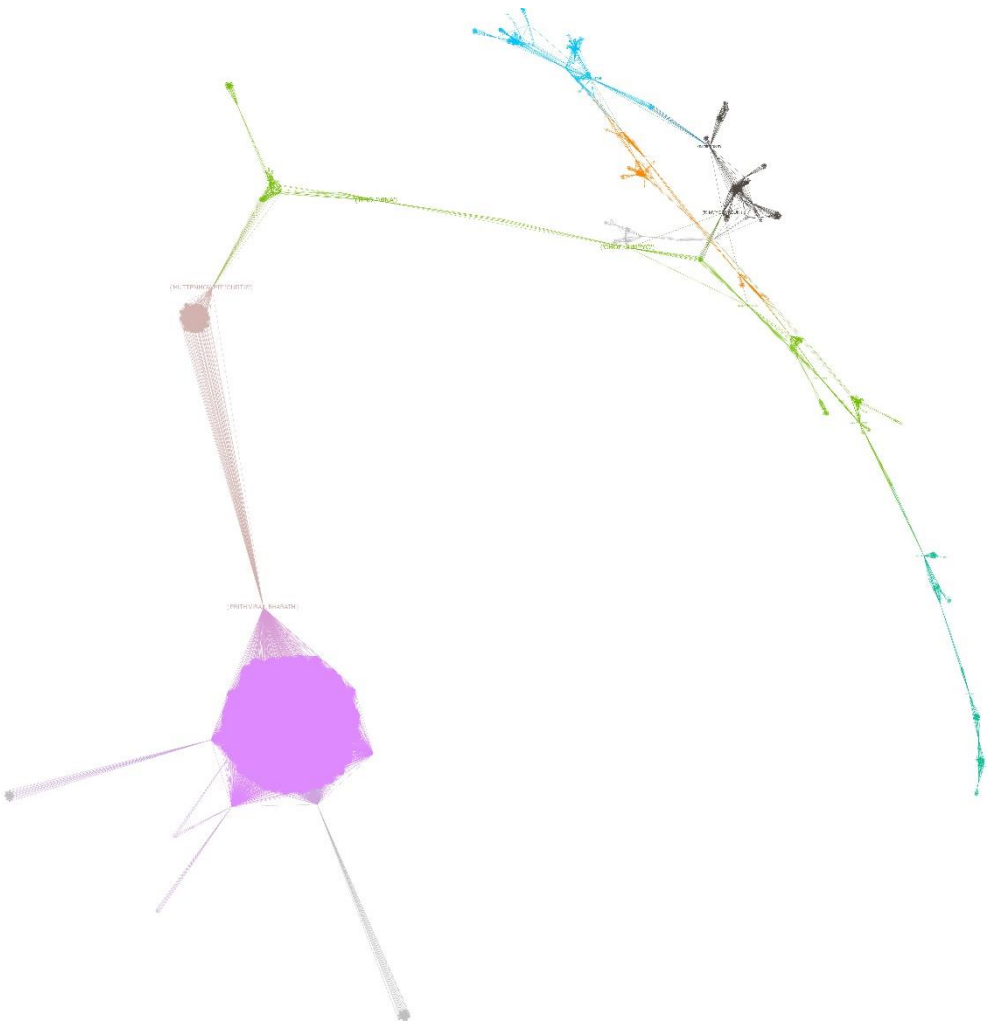

Figure III.4.29. The microbiome Co-Author Networks in Korea Early Phase

Ratio of node and edges: 1090 node (44.4%) ,51732 edges (89.38%). Top Highest Betweenness Author: ('RHO', 'MINA') 290356.2 ('HUTTENHOWER', 'CURTIS') 279968 ('PRITHIVIRAJ', 'BHARATH')260948 ('KIM', 'YOONKEUN') 213814.2738 ('IM', 'SINHYEOG') 160455.82

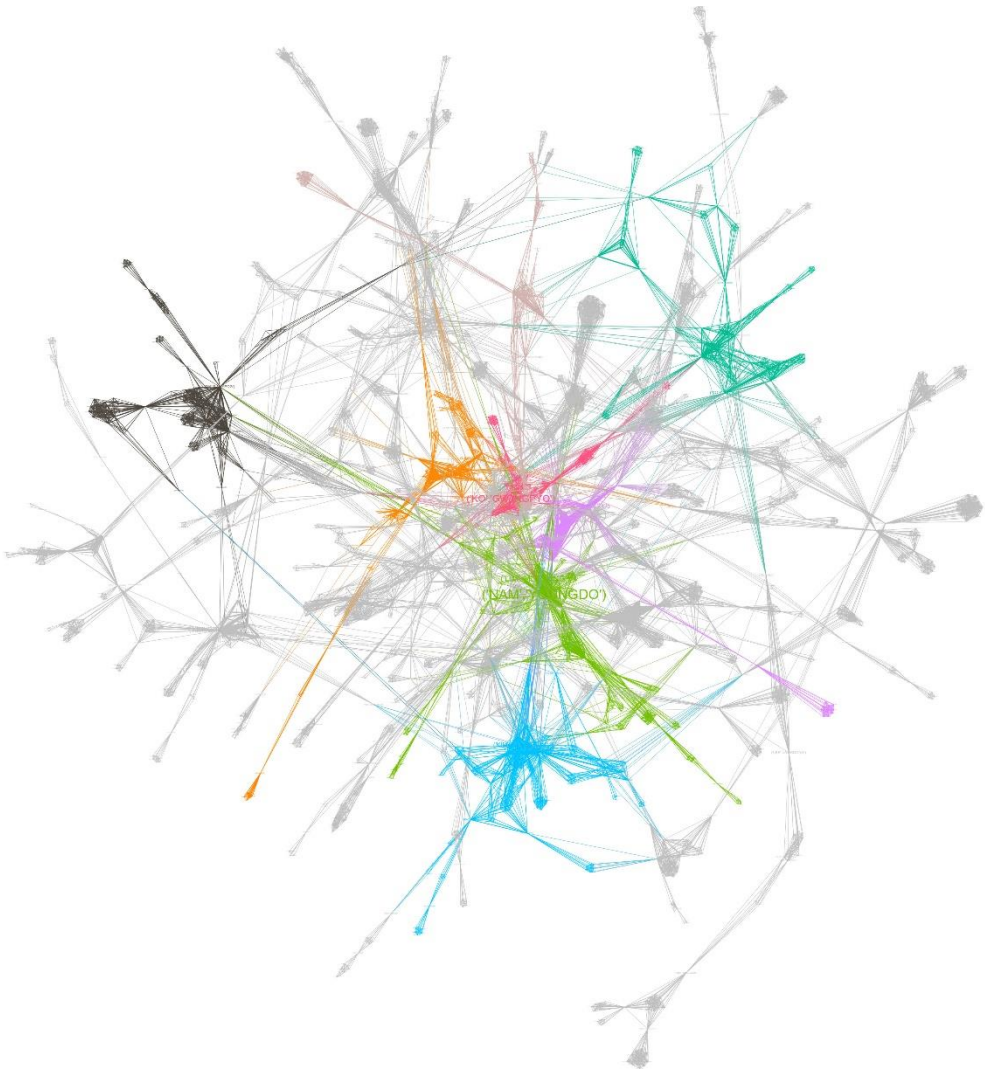

Figure III.4.30. The microbiome Co-Author Networks in Korea Mid Phase

Ratio of node and edges: 3114 node (66.11%) ,18247 edges (60.69%). Top Highest Betweenness Author : ('NAM', 'YOUNGDO') 575370.1735 ('KO', 'GWANGPYO') 450193.9423 ('LIM', 'MI YOUNG') 403524.2672 ('KIM', 'DONGHYUN') 332088.1772 ('KIM', 'YOONKEUN') 324753.911

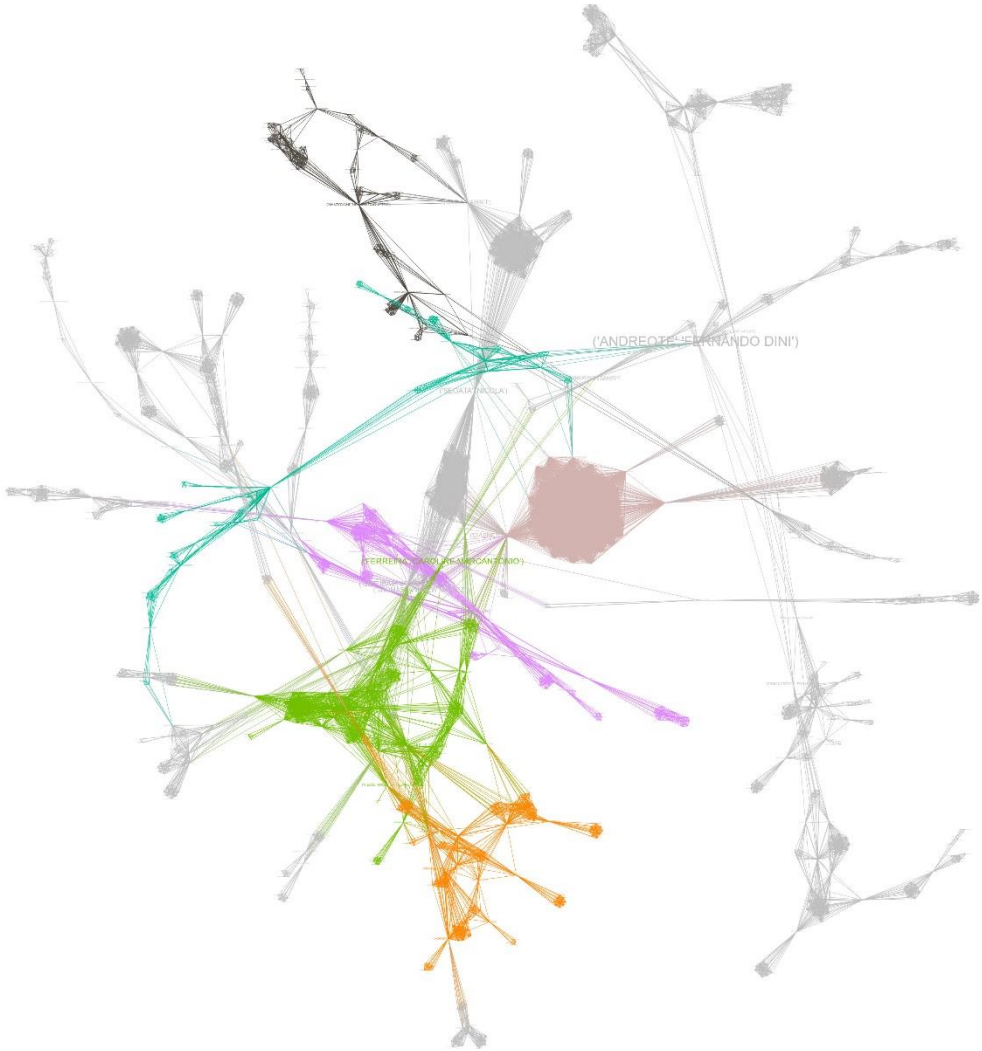

Figure III.4.31. The microbiome Co-Author Networks in Brazil Early Phase

Ratio of node and edges: 1548 node (32.48%) , 61900 edges (81.21%). Top Highest Betweenness Author : ('PRITHIVIRAJ', 'BHARATH') 585646 ('DIASNETO', 'EMMANUEL') 572005.1961 ('SELDIN', 'LUCY') 388080 ('VOLLU', 'RENATA ESTEBANEZ') 384396 ('SALMON', 'DIDIER') 357712 ('MARTINS', 'FLAVIANO DOS S') 307134.3603

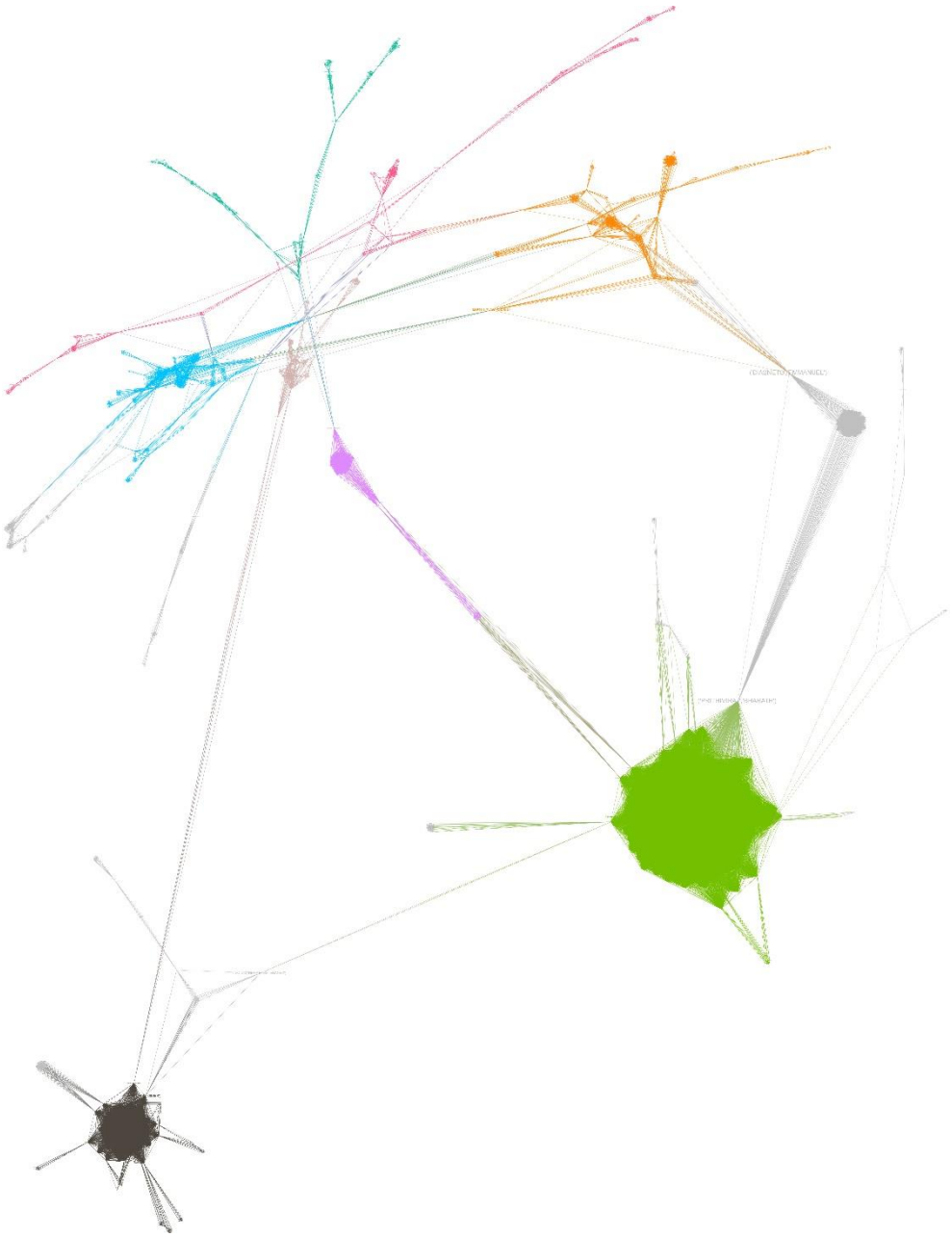

Figure III.4.32. The microbiome Co-Author Networks in Brazil Mid Phase

Ratio of node and edges: 1844 node (25.94%) ,18919 edges (34.17%),. Top Highest Betweenness Author : ('ANDREOTE', 'FERNANDO DINI') 603310.1222, ('FERREIRA', 'CAROLINE MARCANTONIO') 414996.5015 ('SETUBAL', 'JOAO CARLOS') 385845.4581 ('SEGATA', 'NICOLA') 381758.6923 ('THOMAS', 'ANDREW MALTEZ') 349036.6958 ('DIASNETO', 'EMMANUEL') 324745.1265

Chapter IV. DISCUSSION

IV.1. Advancement of research network structure

We closely examined that the network of scientific researchers approximates the characteristics of a 'Small World Network', rather than a random model. However, further investigation is needed to examine whether the distribution of nodes follows a 'Scale-Free Network' structure. A Scale-Free Network, where the distribution of top nodes is represented as a power law, often refers to these key nodes as hubs.

Determining the structure of network is meaningful for several reasons. Firstly, it allows us to validate if the scientific network follows Pareto's principle, where a majority of co-work is conducted by top-tier modules. Secondly, in a Scale-Free Network, the concept of preferential attachment exists in network growth and hub creation. Applying this to a scientific network, if it strongly exhibits Scale-Free characteristics, it implies that top researchers existed from the very beginning with preferential attachment, and this preferential attachment greatly influences the subsequent development of the scientific network.

Understanding this preferential attachment would greatly aid in comprehending the nature of the scientific network. If the network leans more towards a Small World Network structure than a Scale-Free Network, this also holds significant implications. Based on our analysis so far, nodes exhibiting preferential attachment exist, and if these nodes show considerable dynamics, the node distribution can be expressed exponentially, thereby describing the network through persistent dynamic network structural changes. The crucial concepts in a Scale-Free Network, namely network growth and the creation of hubs with preferential attachment, are present in a Small World Network, albeit with node ratios represented by an exponential function rather than a power law.

Elucidating the nature of these hub networks is critical, as they can determine the characteristics of the network. Stated differently, this forms a substantial basis for explaining the factors of how a scientific network evolves. Specifically, if we can

elucidate the factors behind the degree of hub preferential attachment, its emergence, subsequent increase, and eventual decrease, we can construct a compelling narrative to explain scientific progress.

In this study, we looked at researchers closely related to technology as hubs showing preferential attachment. However, this alone cannot fully explain the factors of network growth. There are other potential factors that we aim to discuss, and these discussions are very important for future research. One aspect that could be indirectly observed in this study was the potential impact of diseases or research institutions on the network. Consortiums and research grants are also very important factors for explaining the preferential attachment of hubs. These factors are still a subject of discussion and are likely to be significant research targets in the future. Understanding these variations in the average clustering coefficient provides insights into the changing nature of scientific collaboration and network structure within each country's microbiome research community.

IV.2. Impact of sequencing technology

With the advancements in DNA sequencing technology, it has become possible to analyze human microbiome genes with a certain level of accuracy. Current research estimates that the human microbiome consists of approximately 500-1,000 bacterial species, and if each species has around 2,000 genes, the total number of genes in the human microbiome is estimated to be 2 million (Heather and Chain, 2016). Analyzing research trends at the gene level is a meaningful approach in life science research. Manipulating the genes of cultivable microorganisms is a crucial research method in human microbiome research, as the vast majority of microorganisms cannot be cultured in the laboratory. Therefore, in this study, our aim is to analyze the names and functions of genes used in human microbiome research as the central criteria for analysis.

In the field of microbiome research, technological advancements can be further categorized. Initially, researchers faced limitations in directly cultivating a large number of microorganisms in the laboratory. Therefore, many studies have progressed alongside the development of DNA sequencing methods. One such method is pyrosequencing, which was developed in 2004. Pyrosequencing significantly reduced the time and cost required for analysis (30 Mb/h, $10/Mb), enabling bacterial whole-genome sequencing and facilitating small-scale studies of 16S rRNA community analysis even in smaller laboratories. The 16S rRNA analysis allows for the examination of the entire community by analyzing only about 500 base pairs per species using universal primers, instead of analyzing the entire 16S rRNA (~1,550 base pairs). (Table IV.2.1)

Technologies developed after 2006, such as the Genome analyzer, further reduced the time and cost required for analysis (10 Gb/h, $0.07/Mb), thereby expanding the scope of analysis to include gene expression analysis, such as RNA sequencing (RNA seq). This expansion of technology broadened the range of research fields that could be explored.

Overall, these advancements in DNA sequencing technology have revolutionized microbiome research by enabling more comprehensive analysis of genes and gene expression, and by providing insights into the complex composition and functions of the human microbiome.

But in this study, introduction period of sequencing technology (ex : sequencer) does not vary significantly between countries. However, there were differences between countries in researcher nodes that performed methodological research to analyze and interpret the sequencing results. Notably, several top nodes seemed to function like hub nodes.

Table IV.2.1. Comparison of device performance for representative sequencing technologies

Sequencer	**Sanger 3730xl**	**454 GS FLX**	**HiSeq 2000**
Sequencing mechanism	Dideoxy chain termination	Pyrosequencing	Sequencing by synthesis
Detection	laser	CCD	CCD
Amplification approach	PCR	Emulsion PCR	Bridge amplification
multiple approach	micro capillary	Picotitre wells with microbeads	solid-phase
Output data/run	1.9~84 Kb	0.7 Gb	600 Gb
Cost/million bases	$2400	$10	$0.07

IV.3. Diseases and microbiome networks

The first point highlighted in the network analysis relates to the emergence of top nodes, i.e., researchers influenced by technological advancement. However, it could be postulated that research on diseases such as diabetes, IBD, and obesity would occupy a significant proportion of the network. Furthermore, the approach to these diseases might vary from country to country. Especially in the case of China, Lin Juan, a top node emerging in the mid-phase, is identified as a researcher related to lung diseases. In contrast, it was challenging to identify researchers of similar characteristics within the top nodes in other countries. This disparity could suggest another form of preferential attachment influencing network growth and structure related to interest in diseases. Although, a hurdle in exploring this aspect lies in the need to match each paper with specific diseases, overcoming this challenge to examine the network from a disease perspective could hold significant meaning.

IV.4. Institutions and microbiome networks

Institutions such as universities, research institutes, companies, and hospitals each possess distinct characteristics. Universities carry the aspect of nurturing future scientists. Research institutes can conduct long-term and focused studies on specialized research topics, drawing funds from various sources. Hospitals are research institutions that can most directly approach disease investigation and clinical treatment. To explore the relationship between institutions and network development, we can construct an institutional network using the institutional data provided in WoS. Through this, we can also find out which institutions have prioritized development in the national network. Of note, we noticed that nodes associated with BGI were found among the top nodes in the Chinese network. Providing explanations for such aspects would be meaningful.

IV.5. Consortia and research funding

Consortia, collaborative partnerships between scientific institutions, drive scientific advancement by pooling expertise, resources, and funding. Research funding variations across countries impact scientific development. International collaborations facilitate global knowledge exchange and innovation. Network analysis helps identify and analyze consortia, mapping collaborative relationships and revealing collaboration patterns. It also aids in understanding funding networks, comparing consortia and funding mechanisms to gain insights into their effectiveness. Ultimately, these factors shape the progress of science and knowledge advancement.

For that Network analysis can help identify and analyze consortia by examining collaborative relationships between scientific institutions. By mapping the connections and collaborations among institutions, researchers can identify clusters or groups of institutions that work together in specific research areas. Network metrics such as centrality, density, and community detection algorithms can provide information about the structure and characteristics of consortia. Network analysis can reveal collaboration patterns within and between consortia. Researchers can analyze the strength and frequency of collaborations between institutions within a consortium and identify key players or influential institutions based on centrality measures such as degree centrality or betweenness centrality. By studying the collaboration patterns, researchers can gain insights into the dynamics of knowledge sharing, resource exchange, and expertise distribution within consortia. Network analysis can also be applied to understand the variations in research funding across countries. Researchers can examine the relationships between funding agencies, institutions, and individual researchers to map the flow of funding. This can involve tracking grants, funding allocations, and collaborations between funding agencies and research institutions. By analyzing the funding networks, researchers can identify patterns of funding distribution, dominant funding sources, and potential gaps in research funding within and across countries. Network analysis allows for comparative studies between different consortia or countries. Researchers can compare the structural properties of consortia networks, such as network size, density,

and connectivity, to identify similarities and differences. Additionally, researchers can compare funding networks to understand variations in research investment, funding sources, and the impact of funding on scientific collaborations and outcomes. These comparative analyses can provide valuable insights into the factors that contribute to the success and effectiveness of consortia and research funding mechanisms.

IV.6. Two mode networks

The development of network measures is crucial to facilitate accurate quantitative comparisons between different networks, particularly those with varying numbers of nodes. Currently, limitations exist in conducting precise quantitative comparisons based on betweenness centrality and ACC values across networks with different node counts. It would be more meaningful to have the ability to accurately compare networks from different countries or networks at different time periods.

While different types of networks, such as keyword networks, citation networks, and keyword-author networks, may exhibit distinct characteristics, exploring these alternative networks is challenging due to computational limitations caused by the large numbers of nodes and edges in predicted networks since the mid-2010s. However, as computational power continues to advance, it is worth pursuing such analyses. Additionally, as seen in broader research fields where analyses involving larger data than the current study have yielded impactful results, it is worthwhile to explore these avenues as computational capabilities improve.

Establishing a dedicated data library for biology holds the potential to enhance the quality of topic modeling. By training algorithms used in topic modeling with specialized biological terminology, it becomes possible to incorporate domain-specific knowledge into the analysis. This approach can lead to improvements in the effectiveness and accuracy of topic modeling techniques.

A social network analysis conducted on a plant collection and classification

study in Brazil demonstrated the social constructivist nature of biodiversity research. Through SNA analysis, it was shown that researchers were influenced by senior researchers in their selection of plant collections and could exhibit bias. Additionally, it was found that plant collection involved broader interests rather than solely focusing on researchers' main areas of interest(de Siracusa et al., 2020).

In this study, we start with the assumption that a network exists encompassing all authors who participated in writing research papers, without specifying the directionality of the network. Two-mode networks, also known as bipartite networks or affiliation networks, are used to represent networks where nodes belong to two distinct categories or modes. In such networks, connections exist only between nodes of different categories and not between nodes within the same category. For example, in a movie actor-movie network, the two modes are actors and movies, and the connections represent which actors appeared in which movies.

Two-mode networks possess unique characteristics and analytical techniques that differ from one-mode networks. Bipartite clustering is a common metric used to analyze two-mode networks, measuring the degree to which nodes in one mode are connected to nodes in the other mode. Other common metrics include projection, which transforms a two-mode network into a one-mode network by collapsing one mode, and block modeling, which detects communities in two-mode networks.

Two-mode networks find applications in various fields such as social network analysis, ecology, and information retrieval. In ecology, they represent species-habitat interactions, while in information retrieval, they model user-document relationships in recommender systems.

In our study, we analyzed a two-mode network of authors and keywords in microbiome research to explore collaborative patterns and research topics. We employed bipartite clustering and projection techniques to identify prominent research topics and influential authors in the network. Our findings provide insights into collaborative structures and knowledge dissemination in microbiome research, guiding future research directions in the field.

REFERENCES

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. Proceedings of the International AAAI Conference on Web and Social Media 3, 361–362.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. J. Stat. Mech. 2008, P10008.

Brandes, U., 2001. A faster algorithm for betweenness centrality. The Journal of Mathematical Sociology 25, 163–177.

Brooks, H., 1994. The relationship between science and technology. Research Policy, Special Issue in Honor of Nathan Rosenberg 23, 477–486.

Cullen, C.M., Aneja, K.K., Beyhan, S., Cho, C.E., Woloszynek, S., Convertino, M., McCoy, S.J., Zhang, Y., Anderson, M.Z., Alvarez-Ponce, D., Smirnova, E., Karstens, L., Dorrestein, P.C., Li, H., Sen Gupta, A., Cheung, K., Powers, J.G., Zhao, Z., Rosen, G.L., 2020. Emerging Priorities for Microbiome Research. Front Microbiol 11, 136.

Daru, B.H., Park, D.S., Primack, R.B., Willis, C.G., Barrington, D.S., Whitfeld, T.J.S., Seidler, T.G., Sweeney, P.W., Foster, D.R., Ellison, A.M., Davis, C.C., 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. New Phytol 217, 939–955.

de Siracusa, P.C., Gadelha, L.M.R., Ziviani, A., 2020. New perspectives on analysing data from biological collections based on social network analytics. Sci Rep 10, 3358.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Presented at the NAACL-HLT 2019, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.

Dolgin, E., 2017. The most popular genes in the human genome. Nature 551, 427–431.

Freeman, L.C., 1977. A Set of Measures of Centrality Based on Betweenness.

Sociometry 40, 35–41.

GNU, P., 2007. Unix shell program.

Guido, van R., 1995. Python Software Foundation.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362.

Heather, J.M., Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1–8.

Kemp, C., 2015. Museums: The endangered dead. Nature 518, 292–294.

Lambiotte, R., Delvenne, J.-C., Barahona, M., 2014. Laplacian Dynamics and Multiscale Modular Structure in Networks. IEEE Trans. Netw. Sci. Eng. 1, 76–90.

Latapy, M., 2008. Main-memory triangle computations for very large (sparse (power-law)) graphs. Theoretical Computer Science 407, 458–473.

Lederberg, J., McCray, A., 2001. `Ome Sweet `Omics--A Genealogical Treasury of Words. The Scientist.

Marchesi, J.R., Ravel, J., 2015. The vocabulary of microbiome research: a proposal. Microbiome 3, 31.

McKinney, W., 2010. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference 56–61.

Milestones in Human Microbiota Research [WWW Document], n.d. URL (accessed 4.18.23).

Newman, M.E.J., 2004. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences 101, 5200–5205.

Newman, M.E.J., 2001. The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences 98, 404–409.

OpenAI, 2023. ChatGPT.

Park, J.-M., Lee, J.-H., Hong, S.-I., 2014. Analysis of Research Trend in Human Microbiome Food science and industry 47, 80–91.

Peirson, E., 2016. Tethne.

Prescott, S.L., 2017. History of medicine: Origin of the term microbiome and why it matters. Human Microbiome Journal 4, 24–25.

Scott, J., 2012. Social Network Analysis, 3rd edition. ed. SAGE Publications Ltd, Los Angeles.

Smith, B., 1988. Foundations of Gestalt Theory [WWW Document]. URL https://philarchive.org/rec/SMIFOG (accessed 4.19.23).

Stanley, M., 1967. The Small World Problem. Psychology Today, DJJR Sociology Vol. 2, 60–67.

Stella, P., Rem, K., 2017. Scientometric Analysis of Human Microbiome Project. Journal of Siberian Federal University Humanities & Social Sciences 10, 1076–1082.

Structure, function and diversity of the healthy human microbiome | Nature [WWW Document], n.d. URL https://www.nature.com/articles/nature11234 (accessed 5.31.23).

Waldor, M.K., Tyson, G., Borenstein, E., Ochman, H., Moeller, A., Finlay, B.B., Kong, H.H., Gordon, J.I., Nelson, K.E., Dabbagh, K., Smith, H., 2015. Where Next for Microbiome Research? PLoS Biol 13, e1002050.

Wang, J., Kuenzel, S., Baines, J.F., 2014. Draft Genome Sequences of 11 Staphylococcus epidermidis Strains Isolated from Wild Mouse Species. Genome Announc 2, e01148-13.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of ‘small-world’ networks. Nature 393, 440–442.

Wei, C., Brent, M.R., 2006. Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 7, 327.

국문 초록

우리는 이번 연구에서 과학 연구 성장의 구조와 원인을 찾기 위하여 서지학적인 분석, 네트워크 분석을 수행하였다. 최근에 급격하게 발전하고 연구 규모가 거대한 마이크로바이옴 연구 분야를 선택하여 분석하였다. 특히 이번 연구에서 전세계 마이크로바이옴 연구 규모 상위 11개 국가의 분석을 수행하였는데, 해당 국가들의 네트워크를 2000년부터 2021년에 걸쳐 연도별 변화를 분석하였다. 이를 통해 국가별 연구 네트워크 변화의 공통점과 차이점을 확인할 수 있었다.

서지학 분석으로 바라본 성장 패턴에서 공통적으로 과학 네트워크의 성장을 예측할 수 있었다. 서지학적 접근을 더 자세히 분석하기 위해 공저자 네트워크를 제작하였다. 대부분 국가의 네트워크가 서지학 분석을 통해 예측한 것보다 성장에 차이가 없을 것이라고 분석하였다.

이 후 네트워크 이론에서 사용되는 여러가지 측정값들을 사용하여 네트워크의 구조를 표현하였다. 우선 인접 노드 간의 결집도를 정량화 하는 평균 집결 계수를 살펴보았다. 그 결과 과학 연구분야는 초기부터 높은 값을 가진 채로 네트워크가 생성되고 시간이 지나면서 조금씩 값이 감소하는 모습을 보였다. 이것은 네트워크를 이루는 연구자들이 전체 성장 시기에 걸쳐 매우 단단히 연결되어 있고, 시간이 지나면서 그 연결성이 조금씩 떨어지는 것을 의미한다. 그런데 중국 네트워크에서는 다른 국가들에 비해 훨씬 빠르게 평균 집결 계수가 감소하는 것을 볼 수 있었다.

또 네트워크의 정보 효율성 등을 나타내는 평균 경로길이의 변화가 독특함을 알 수 있었다. 평균 경로길이의 증가는 네트워크 크기에 증가에 비례한다고 알려져 있다. 그런데 대부분의 네트워크에서 시그모이드 함수(Sigmoid function)의 모양으로 증가하고 일정한 값에 수렴하였고, 특정 국가들은 선형 함수 (Linear fuction)의 모습으로 증가하고 일정한 값에 수렴하였다. 네트워크가 생성되고 발전하는 동안

평균 경고길이가 어떤 시점에서 특정 값으로 수렴한다는 점은 지금까지 알려지지 않았던 내용이다.

이번 연구는 랜덤 그래프 생성 모델인 ‘에르되스 레니 모델’(Erdős - Rényi model)이나 ‘작은 세계 네트워크’(Small world network) 생성 모델인 ‘와츠-스트로가츠 모델’(Watts-Strogatz model)과 같은 네트워크 생성 모델링을 실제 네트워크 사례에서 보여주었다.

해당 연구 결과에서는 마이크로바이옴 연구분야가 ‘와츠-스트로가츠 모델’에 의해 생성되는 높은 평균결집계수와 짧은 평균 경로길이를 가지는 '작은 세계 네트워크’에 근사할 것으로 생각한다. 특히 생성되는 시점에서도 '작은 세계 네트워크’에 근사하며 크기가 커지면서도 "작은 세계 네트워크’에 근사한 채로 커지는 것도 알 수 있었다. 평균결집계수가 점차 감소하는 것을 본다면 ‘'작은 세계 네트워크’에 근사함이 지속적으로 유지되지 않을 수도 있음을 보여주었다.

이러한 구조적인 변화속에서 네트워크 성장의 요인, 과학 성장의 요인을 규명하기 위하여 각 네트워크의 주요 노드들의 성격을 살펴보았다.
서로 다른 연구자 사이에 해당 노드가 얼마나 존재하는지 수를 의미하는 사이 중심성을 통해 살펴보았다.
각 네트워크의 주요 노드들의 지위는 네트워크의 발전이 진행되면서 일반적으로 동적인 변화를 보였다. 특히 네트워크의 최상위 노드들도 네트워크 내의 연결성의 지위를 분석 기간동안 똑같이 유지하지는 않았다. 이것은 과학 네트워크의 발전이 어느정도 동적이라는 것을 예측하게 해준다. 그리고 미국과 중국의 경우 주요 노드들 중 염기서열 분석 기술과 관련되어 있는 연구를 하는 노드들이 각 네트워크에서 지위의 상승 및 허브 노드로 작용하였다. 또 이 노드들은 타 국가 네트워크에 끼어들기를 함을 확인할 수 있었다.
이를 통해 네트워크의 발전 요인중에 염기서열 분석에 대한 연구자들의 접근이 크게 영향을 준 것을 알 수 있었다..

이러한 발견들은 마이크로바이옴 분야, 더 나아가 과학이 발전하는

방식을 설명해 줄 수 있다는 점에서 본 연구의 의미가 있다.

Keywords: 마이크로바이옴 연구, 네트워크 분석, 공저자 네트워크, 고성능 염기서열분석법, 기술 발전