NETWORK ANALYSIS OF SCIENTIFIC COLLABORATION AND CO-AUTHORSHIP OF THE TRIFECTA OF MALARIA, TUBERCULOSIS AND HIV/AIDS IN BENIN.

by

Gbedegnon Roseric Azondekon

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin − Milwaukee

August 2018

ABSTRACT

NETWORK ANALYSIS OF SCIENTIFIC COLLABORATION AND CO-AUTHORSHIP OF THE TRIFECTA OF MALARIA, TUBERCULOSIS AND HIV/AIDS IN BENIN.

by

Gbedegnon Roseric Azondekon

The University of Wisconsin-Milwaukee, 2018
Under the Supervision of Professor Susan McRoy

Despite the international mobilization and increase in research funding, Malaria, Tuberculosis and HIV/AIDS are three infectious diseases that have claimed more lives in sub Saharan Africa than any other place in the World. Consortia, research network and research centers both in Africa and around the world team up in a multidisciplinary and transdisciplinary approach to boost efforts to curb these diseases. Despite the progress in research, very little is known about the dynamics of research collaboration in the fight of these Infectious Diseases in Africa resulting in a lack of information on the relationship between African research collaborators. This dissertation addresses the problem by documenting, describing and analyzing the scientific collaboration and co-authorship network of Malaria, Tuberculosis and HIV/AIDS in the Republic of Benin.

We collected published scientific records from the Web Of Science over the last 20 years (From January 1996 to December 2016). We parsed the records and constructed the coauthorship networks for each disease. Authors in the networks were represented by vertices and an edge was created between any two authors whenever they coauthor a document together. We conducted a descriptive social network analysis of the networks, then used

mathematical models to characterize them. We further modeled the complexity of the structure of each network, the interactions between researchers, and built predictive models for the establishment of future collaboration ties. Furthermore, we implemented the models in a shiny-based application for co-authorship network visualization and scientific collaboration link prediction tool which we named **AuthorVis**.

Our findings suggest that each one of the collaborative research networks of Malaria, HIV/AIDS and TB has a complex structure and the mechanism underlying their formation is not random. All collaboration networks proved vulnerable to structural weaknesses. In the Malaria coauthorship network, we found an overwhelming dominance of regional and international contributors who tend to collaborate among themselves. We also observed a tendency of transnational collaboration to occur via long tenure authors. We also find that TB research in Benin is a low research productivity area. We modeled the structure of each network with an overall performance accuracy of 79.9%, 89.9%, and 93.7% for respectively the malaria, HIV/AIDS, and TB coauthorship network.

Our research is relevant for the funding agencies operating and the national control programs of those three diseases in Benin (the National Malaria Control Program, the National AIDS Control Program and the National Tuberculosis Control Program).

To my family...

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AIC	**A**kaike's **I**nformation **C**riterion
AIDS	**A**cquired **I**mmune **D**eficiency **S**yndrome
AND	**A**uthor **N**ame **D**isambiguation
ARV	**A**ntiretroviral
AUC	**A**rea **U**nder the **C**urve
AWS	**A**mazon **W**eb **S**ervices
BIC	**B**ayesian **I**nformation **C**riterion
CD4	**C**luster of **D**ifferentiation **4**
CGI	**C**ommon **G**ateway **I**nterface
CI	**C**onfidence **I**nterval
CPU	**C**entral **P**rocessing **U**nit
df	**d**egree of **f**reedom
DOI	**D**igital **O**bject **I**dentifier
ELISA	**E**nzyme-**L**inked **I**mmunosorbent **A**ssay
ERGM	**E**xponential **R**andom **G**raph **M**odel

Fig.	**F**igure
GLM	**G**eneralized **L**inear **M**odel
HIV	**H**uman **I**mmunodeficiency **V**irus
HTTP	**H**yper**T**ext **T**ransfer **P**rotocol
ICL	**I**ntegration **C**lassification **L**ikelihood
JSON	**J**ava**S**cript **O**bject **N**otation
LNM	**L**atent **N**etwork **M**odel
MeSH	**Me**dical **S**ubject **H**eadings
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MCMLE	**M**onte **C**arlo **M**aximum **L**ikelihood **E**stimation
MDG6	**M**illenium **D**evelopment **G**oal **6**
MLE	**M**aximum **L**ikelihood **E**stimation
MPLE	**M**aximum **P**seudo**L**ikelihood **E**stimation
ORCID	**O**pen **R**esearcher & **C**ontributor **ID**
PA	**P**referential **A**ttachment
PR	**P**recision **R**ecall
REF	**R**eference
ROC	**R**eceiver **O**perating **C**haracteristics
SAOM	**S**tochastic **A**ctor-**O**riented **M**odel
SBM	**S**tochastic **B**lock **M**odel
SCI	**S**cience **C**itation **I**ndex
SCIE	**S**cience **C**itation **I**ndex **E**xpanded

SE	**S**tandard **E**rror
SNA	**S**ocial **N**etwork **A**nalysis
SVG	**S**calable **V**ector **G**raphics
SW	**S**mall **W**orld
TB	**T**uberculosis
TERGM	**T**emporal **E**xponential **R**andom **G**raph **M**odel
URL	**U**niform **R**esource **L**ocator
US	**U**nited **S**tates
WHO	**W**orld **H**ealth **O**rganization
WOS	**W**orld **O**f **S**cience

ACKNOWLEDGEMENTS

First, I would like to express my sincere gratitude to my advisor Prof. Susan McRoy for the continuous support of my Ph.D study, for her patience, motivation, and immense knowledge. Her guidance helped me in my research and writing of this dissertation.

I would also like to offer my special thanks to Dr Charles Welzig and the Welzig Neuroscience and Neurotechnology lab at the Medical College of Wisconsin. I benefited from the computational resource available in the lab to run the analyses and the computationally intensive simulations. My thanks also go to Zachary James Harper for his availability. Without his precious support it would have been impossible to conduct this research.

My special thanks are extended to Dr Spencer (Chiang Ching) Huang of the Joseph Zilber School of Public Health. He has been crucial to the successful continuation of my PhD journey at the University of Wisconsin Milwaukee. The research assistantship he provided me has been a determining factor to my retention in the Doctoral program of Biomedical and Health Informatics.

Finally, I would like to thank the rest of my dissertation committee: Prof Christine Cheng, Dr. Rohit Kate, and Dr. Zhang Qing, for their insightful comments, encouragement, and useful recommendations and suggestions which widen my research from various perspectives.

Chapter 1

General Introduction

Infectious diseases have long claimed the lives of millions of people worldwide. They disproportionately affect the developing nations where 90% of the deaths are caused by very few diseases including Malaria, Tuberculosis (TB) and HIV/AIDS [1]. Malaria, TB and HIV/AIDS remain the three major public health concerns in Sub Saharan Africa where they are responsible for high mortality, morbidity rates and impact negatively on the socioeconomic way of life of the populations [2, 3]. These three diseases have been given special attention at the Millenium Declaration in its 6^{th} Goal of Millenium Development [4]. Initiatives such as the US President's Malaria Initiative, the Global Fund for Malaria, TB and HIV/AIDS and the President's Emergency Plan for AIDS have led to the investment of more than 70 million US dollars to encourage Research and Development, Private-Public partnership as well as to reinforce the activities of non-governmental organizations within the healthcare systems of the affected countries [5–7].

The Global Fund disbursement in 2010 peaked at over 1.45 billion dollars for HIV/AIDS, 416 million dollars for TB and 714 million dollars for Malaria [8, 9]. With these financial supports at hand, efforts have led to a sharp increase of public health interventions and many positive public health outcomes in terms of the reduction of mortality and morbidity related to those diseases [10]. For example, in Benin, the new financing provided for improved entomological surveillance to reduce the morbidity and mortality related to malaria by 75% by 2015. Encouraged and motivated by the success stories in controlling these diseases, some authors formulated the ambitious zero incidence goal of TB and HIV and the zero death goal of the three diseases by 2015 [12].

After the declaration of the Millenium Development Goal 6 (MDG6) in 2000, significant progress has been made in the treatment and prevention of Malaria, TB and HIV/AIDS, leading to the reverse of the mortality and morbidity due to these three diseases. Nevertheless, sub Saharan Africa still carries the burden of these diseases. For example, in 2009, 2.6 million new cases and 1.8 million of death related to HIV were estimated out of which 68% and 72% of respectively new cases and deaths were in Africa [13]. TB cases were estimated at 9.4 million and 1.3 million deaths out of which HIV-positive cases make up 12% of all cases and 23% of all TB deaths [14]. Although the rapid expansion of vector control strategies worldwide, malaria was responsible of 225 million cases and 781,000 death in 2009 out of which over 90% were in Africa [9].

In the Republic of Benin, TB and HIV/AIDS have become a common aspect of the public health system. The three are the main impediments of economic and social progress that are characteristics of poverty. According to a 2000 World Health Organization (WHO)

press report, malaria slows economic growth on the African continent by 1.3% each year [15]. And it is known that Tuberculosis and HIV/AIDS patients experienced severe economic burden in terms of access to health care, treatment and diagnosis [16]. The situation is further compounded by the poorly developed immunity among the children and the elderly, and the predominant malnutrition problem experienced by a majority of the population [17]. Between 2000 and 2013, the impact of the increase in funding has led to an annual decrease in the incidence of 7.6%, 0.6% and 5.2% respectively in HIV/AIDS, TB and Malaria. Similar results were obtained in terms of prevalence with a decrease of 1.3% in HIV/AIDS and 0.8% in TB. Annual death rates decreased also at about 3.1%, 1.2% and 5.3% respectively in HIV/AIDS, TB and Malaria [9, 13, 14].

Successful scientific collaborations have led to the eradication of chickenpox and the near eradication of poliomyelitis through the development of vaccines [17]. For Malaria and HIV/AIDS, the development of a vaccine has proven significantly difficult to develop despite the decades of active research that has not been successful so far [18–20]. This is why researchers need to form continuous and sustainable collaborations through intensive network practices that go beyond the regional boundaries [21]. Scientific collaborations give researchers the opportunity to work and learn from each other. Such collaborations are further needed to overcome the overgrowing challenge of co-infections of HIV/AIDS and Tuberculosis [22, 23].

Despite the increasing financing effort and increasing number of published reports, the literature does not provide sufficient data regarding co-authorship networks of scientific research collaborations and their dynamics in the fields of malaria and TB and HIV/AIDS

research in Africa, and particularly in Benin. This situation results in a lack of information on the main players and drivers of the progress made. As for the eradication of chickenpox [17], collaborative research will undoubtedly play an important role in the successful attainment of the MDG6 in Subsaharan Africa in general and particularly, in Benin. Understanding the structure of these networks is capital since it can help improve research prioritization [24], identify prolific researchers, better design, strategic planning and implementation of research programs [25], and promote cooperation and translational research initiatives [26]. In this dissertation, we document, describe, analyze, and model the different aspects and processes of scientific research collaboration of the three leading infectious diseases in the Republic of Benin. The social network analysis of research collaboration approach is chosen to reveal undiscovered knowledge on effort of researchers in working together towards the reduction of the burden of Malaria, TB and HIV/AIDS. Modern times have rendered research and scientific collaborations irreplaceable policy formulations processes. This is because research collaborations form a stable basis for the provision of evidence based information in the formulation of fundamental principles and guidelines for the elaboration of public health strategies, particularly in developing countries like Benin. For this reason, this dissertation focuses on the Network analysis of the scientific collaborations through co-authorship network analysis.

1.1 General and specific Objectives

The purpose of this research is to analyze the structure and dynamics of scientific collaborations and co-authorship in the fields of Malaria, Tuberculosis and HIV/AIDS research areas over the last 20 years in the Republic of Benin. Our results can help improve grant and research resource allocation to funding and help research organizations and national control programs to promote and encourage trans and interdisciplinary research in the country. Additionally, our findings recommend new approaches to support the Beninese national control programs via better strategic planning and implementation of public health policies, research and development. We also propose a prototype of an online research collaboration tool to assist health policy makers and funding organizations to promote research collaboration in the republic of Benin. More specifically, we address the following research questions:

- What is the structure of scientific research collaboration networks in Benin over the last 20 years in Malaria, TB and HIV/AIDS research?
- Who are the most prolific authors, scientific research groups within each field?
- How have transnational research evolved over the last two decades in the Republic of Benin?
- What are the characteristics and the dynamics of the current co-authorship research collaborations in Benin in Malaria, TB and HIV/AIDS research?

This dissertation fills the gap in the current literature, and reveals the role of the collaborative research in the prevailing research networks. Our research meets the following specific objectives:

1. To identify the most productive and prolific scientific research groups and authors within each research area.

2. To document and describe the structure of Malaria, TB, HIV/AIDS co-authorship networks and their characteristics, how they evolve over time in Benin over the last two decades.

3. To unravel the mechanistic phenomenon explaining the formation and trends of these networks over time.

4. To predict and recommend future research collaboration ties in Benin in the three research areas.

5. To develop a prototype of co-authorship visualization and scientific collaboration tool for Malaria, TB and HIV/AIDS research in Benin.

1.2 Hypotheses

We hypothesize that tie formation in each co-authorship network:

- is dependent on observed authors (vertices) characteristics
- is dependent on the concept of distance in latent space, and
- is dependent on collaboration types and/or membership to a certain research community or cluster.

Chapter 2

A review of the related literature on disease research and applications of network analysis

2.1 Brief Overview of Malaria, Tuberculosis and HIV/AIDS

AIDS is a health condition caused by the Human Immunodeficiency Virus (HIV) [27, 28]. HIV infects and attacks the cells that are responsible for the immune system in the body (CD4 cells) that provide protection against infections and illness. The virus infects the human host by making him vulnerable and unable to fight future infections [29]. The virus eventually weakens and kills the CD4 cells resulting in a weak immune system and

vulnerability to diseases. HIV is transmitted through body fluids exchange, and the infection exists in four stages. The first stage is the primary infection stage and lasts within 2 to 4 weeks. It is characterized by flu-like symptoms, and the infected person is highly contagious. The second stage is the asymptomatic stage that may last for about ten years, and the infected person does not display significant symptoms of the infections. The third stage is the symptomatic stage. At this stage, the virus weakens the immune system, and the infected person suffers from both mild and chronic symptoms as the infected person suffers opportunistic diseases. Illnesses like malaria and TB in HIV infected subjects, are experienced in a severe manner. The fourth stage is AIDS; it causes death within two years if left untreated [29, 30].

According to the World Health Organization (WHO) the signs for HIV/AIDS change through the stages of infections as the disease progresses. To determine whether a person is infected, an HIV test needs to be conducted. ELISA method based HIV testing is one of the most common antibody-based testing method characterized by 99% accuracy rate [27]. It is recommended that a HIV negative test result should be confirmed after three months because the immune system can sometimes take up to 12 weeks to develop the tested antibodies [31]. It is however possible to get false negative results during the 12 weeks window period. The antiretroviral (ARV) drug therapy is initiated when the infected person reaches the third or fourth stage of infection to suppress the virus and boost the immune system. Such measures are taken because there is currently, no cure for HIV and the early initiation of the therapy may result in drug resistance [32–34].

Unlike HIV/AIDS, TB is a highly infectious disease that is caused by a bacteria called

Mycobacterium tuberculosis. The disease exists in active and inactive forms. The active form, also known as the open disease causes the infected person to suffer and to be highly infectious. The inactive/latent TB infection is not infectious, and the infected individual does not suffer from the signs and symptoms associated with the active disease. Healthy individuals with latent infection have approximately 10% probability of getting active TB disease over their life. Chances of infection are high in the first two years after the exposure to the bacteria, and in the case where the host develops any form of lung or immune system damage [35, 36]. On the other hand, in HIV infected individuals co-infected with TB, there exists a 10% annual chance of developing active TB [37–39]. Active TB in adults may result from re-infection with a new strain of TB or perhaps a reaction to the latent infection. Consequently, researchers surmise that silica inhalation, HIV infection, and silicosis are responsible for the high risk of TB infection in the working adults' population [38, 40]. TB symptoms are characterized by a chronic cough, night-time fevers, profuse sweating, and significant weight loss within a short time. However, studies show that people with TB can be infectious prior to showing the symptoms or complaining of any form of pulmonary discomfort. In the worst case scenario, TB goes beyond the pulmonary and infects other parts of the body, especially for people infected with HIV. HIV complicates the manifestation of TB in terms of its symptoms and signs in 70% of the HIV/AIDS infected population suffering from TB [38]. Studies indicate that people with undetected open TB disease are the leading cause of TB infections. Even though TB is a treatable disease, the treatment procedure is extremely aggressive. The treatment procedure for first-time patients entails administration of a six-months dose under

close medical supervision termed as directly observed therapy. The other challenge in the treatment is that there are approximately 25% TB-drugs resistance cases worldwide every year [41, 42]. Approximately 80% of people with TB can be cured of their active TB infection, however, HIV and Silicosis increases the risk of reinfection by 20%. The infection among individuals with silicosis, may cumulatively contribute to lung damage and work inability. Additionally, the HIV/AIDS increases the risk of opportunistic infections, which may result in a poor outcome for the TB treatment [43–45].

Completing the trifecta is malaria, a parasitic infectious disease caused by the Plasmodium parasites. Even though, malaria is predominantly found in the tropical regions, 48% of the instances of infections have been experienced in the Northern and Southern parts of America, Asia, and Africa, putting approximately 50% of the world's population at risk. The malaria pathogens are *Plasmodium ovale*, *Plasmodium malariae*, *Plasmodium vivax*, and *Plasmodium falciparum* which is the deadliest. The distribution of the disease matches that of its vectors, the female mosquitoes of the genus Anopheles [46, 47]. In the sub-Saharan African countries, the vector of the disease is *Anopheles gambiae s.l.* Malaria has a range of symptoms and signs that manifest differently from one person to another. The most common symptoms are fevers, gastrointestinal symptoms, and fatigue, headaches, and muscle aches. The malaria pathogen infects two hosts, the Anopheles mosquito, and the infected human. When the infected mosquito feeds from an individual, it injects sporozoites into the circulatory system of the bitten person. The sporozoites reside in the liver cells until they become mature schizonts. The schizonts rupture upon maturity and release merozoites, which infect the red blood cells [48]. The two most used

malaria test are rapid tests using an instant result kit akin to the home pregnancy test device, and the blood smear test that is examined under the microscope for the presence of red blood cells that are infected by the parasite. Treatment entails administration of drugs that range in types. While some malaria drug prescriptions may have a three days dosage, others may have up to one week dosage [49, 50].

HIV/AIDS, TB, and Malaria form together a trifecta of diseases caused respectively by a virus, a bacteria, and a parasite.

2.2 Network Analysis of Scientific Research collaboration

Collaboration in science is essential to research and development, knowledge discovery, technology and innovation. The effectiveness of collaboration in science can be measured using scientometrics. According to Leydesdorff and Milojevic [51], scientometrics uses quantitative and computational methods to analyzing and measuring science, communication in science and science policy. The field of scientometrics emerged from Eugene Garfield's idea to improve Information Retrieval [52], followed by the creation of the Science Citation Index (SCI) in the 1960s, and the availability of scientific databases references publications. The discipline of Scientometrics is aimed at providing guidance to several research issues involving the measurement of science impact, the measurement of impact journals and institutional units, theories of citation, and the mapping of science. Here, we focus on the mapping of science since it is essential to understanding the

dynamic of science, informing policy decisions, and identifying important fields, research groups, as well as specialties based on evidence from the literature [51]. Such goals can be achieved by mapping publications, authors and analyzing patterns of collaborations between them.

Since the publication of the first co-authored paper in 1665, scientific co-authorship has spread significantly throughout the scientific realm and the number of co-authored scientific publications have tremendously increased [53]. According to Wagner [54], the increase in international scientific co-authorship has been of a fast growth. International co-authorship originates from international collaborations between scientists. In general, international collaborations have more visibility than national collaborations and often result in publications in high impact journals [55].

The paradigm of co-authorship network is rooted in network theory. In a co-authorship network, the reasearchers are represented by the set of vertices and the relationship between them are represented by the set of edges. An edge between two researchers in such a network means that they both coauthor a publication. Unlike citation networks, the scientific community has dedicated less attention to co-authorship networks because of the long tradition of citation network analysis in bibliometric [21, 56]. Nevertheless, the analyses of how complex co-authorship networks form and evolve in time is crucial for identifying leading researchers in a particular scientific domain, describing their extant to collaborate with their peers, and evaluating the impact of their research [26]. An example of such an investigation is illustrated in Newman scientific collaboration paper series on Biomedical research, physics and computer science co-authorship networks [21, 56–58].

Taking publications as units, the analyses of scientific collaboration facilitate the study of trans and inter-disciplinary research by focusing on the dynamics of the collaboration networks [59]. In addition, these networks can provide important information regarding cooperation patterns among authors and their status and location in the structures of the scientific community [60]. Furthermore, Mali et al. [61] assert that co-authorship social network studies are highly relevant for funding organizations for promising and emerging topics support in science.

Although many authors have proposed different features for classifying co-authorship networks [62–64], the categorization features of Andrade et al. [62] identifies three levels of classification of scientific collaboration: the cross-disciplinary level with the intradisciplinarity and interdisciplinarity subdimensions, the cross-sectoral level with the intramural and extramural research collaboration subdimensions and the cross-national level including the national and international scientific collaboration subdimensions. For a full description of each level of scientific collaboration, we refer the reader to Mali et al. [61]. The methods of co-authorship network studies have emerged from social network analysis and graph theory. Such studies heavily relied upon access to scientific collaboration data sources such as SCOPUS, the Web Of Science, PubMed, Medline or even Google Scholar. In general, Mali et al. [61] identify three methodological approaches to studying scientific co-authorship networks:

(i) basic analysis of network properties using temporal data (usually in the

form of a time-series of snapshots), (ii) deterministic approaches to the analysis of scientific co-authorship networks, and (iii) statistical modeling of network dynamics

In addition to the three approaches outlined by Mali et al. [61] and mentioned above, co-authorship networks can be analyzed on the basis of formal network properties, including network degree, density, path, path length, shortest path and the global clustering coefficient. Many scientific collaboration network studies have adopted this graph-based approach to scientific co-authorship investigation. In the next paragraphs, we present and discuss the purpose, methods and the results of some of those studies.

Newman [21] investigated scientific network collaboration in biomedical research, physics and computer science. In this study, Newman collected data from four databases and presented distribution of collaboration networks, demonstrated the presence of clustering and highlights differences between the scientific fields under investigation. According to his findings, Newman [21] concluded on the "smallworldness" of such networks in which scientists are only separated by shorter paths. In a second paper published the same year, Newman [57] provided a deeper analysis of the networks using the same data. He presented a variety of statistical properties of the networks, identified giant collaborative components and study centrality and connectedness measures. In Newman [58], the author evaluated various nonlocal network properties including shortest paths and distance between researchers. He proposed a modified version of the standard breadth-first search algorithm for evaluating the geodesic distance between the scientists in the network. He later weighted the networks by the number of paper published by pairs of researchers as

well as their number of coauthors, and calculated all the distances using Dijkstra's algorithm. His analyses provided insights in the strength of collaboration in each network. In a last paper in the same series, the author summarized the results of the three previous studies and showed how patterns of collaboration varied between scientists within a scientific field over time [56].

In another study, Hou et al. [65] applied a variety of graph-based algorithms to quantify the importance and impact of science, analyzing data retrieved from the Science Citation Index (SCI) over a period expanding from 1978 to 2014. In addition to methods of Social Network Analysis (SNA), the authors used co-occurrence analysis, cluster analysis and frequency analysis of words to describe the microstructure of the scientometrics network, revealing the major collaborative clusters and identifying the center of the scientometrics collaborative network. All analyses were performed using a free online software called Bibexcel and visualizations were displayed using the Pajek program. Similarly, to Newman's publications, this paper applied basic network analysis based on network properties such as degree, closeness and betweenness centrality measures. Unlike Newman's studies, it also accounted for citation data. Yet another paper reported the collaborative patterns in co-authorship network in the scientific discipline of reproductive biology [66]. This study conducted a bibliometric analysis on 4,702 papers published in the field from 2003 to 2005. Although their analysis was basic, the study did not make use of any network property measures but was rather, mainly descriptive. Nevertheless, the study identified important components by applying an unspecified clustering algorithm using the Bibliométricos software, and the Pajek program for data visualization. A similar

bibliometric analysis is also reported by Toivanen and Ponomariov [67] who investigated the research collaboration patterns in the African regional systems. Their data were publication records from African institutions from 2005 to 2009, processed via a proprietary text mining software named VantagePoint. Analysis of the network was performed using the UCINET software. The authors adopted an empirical clustering method based on the geographic regions within the African research context. Their research uncovered the dynamic nature of African collaborative efforts despite the lack of research capabilities, the structural weaknesses, and the uneven integration of resources.

Some researchers have studied scientific network co-authorship across a scientific discipline in specific institutions or organizations. For example, Bellanca [68] used basic network analysis to measure interdisciplinary research by describing three co-authorship networks of researchers in Biology and chemistry departments at the University of York. After extracting publication records from the Web Of Science, the author used the Bibexcel tool and the UCINET software to analyze the co-authorship networks. The analysis was descriptive involving the assessment of basic network properties such as node degree, betweenness, and clustering coeeficient. They discovered fewer interdisciplinary research between biologists and chemists within the University but more interdisciplinary links between biology and mathematics, bioinformatics, biophysics and biochemistry. Their findings are potentially important for the development of strategies to promote interdisciplinary research within the University. Another study conducted in a Spanish institution analyzed collaboration between Spanish authors [69]. After retrieving 448 published papers between 1998 and 2007, the authors used basic network analysis, implemented in the

Pajek program, to their network and identify group of authors as well as their relationship with others. In their future directions, the authors recommended that a dynamic time series analysis method as the next step to better understand their co-authorship network. In some other studies, the research focus was on a single country, across a specific scientific discipline.

Using Bibexcel, and the UCINET software package, Ghafouri et al. [24] proposed a sociogram analysis to social co-authorship network of Iranian researchers, in an attempt to help improve research prioritization, research centers establishment, teams and new curricula in the field of emergency medicine. Their results revealed a poorly connected, loose and sparse co-authorship network in the field of emergency medicine in Iran. While their study was keyword based and might have not included all papers, they recommended the rethink of research prioritization, the establishment of new research centers more emergency medicine specialists to Iranian policy makers. Yet another Iranian study by Salamati & Soheili [70] investigated the field of violence, assessing scientific research outputs by Iranian researchers extracted from the Science Citation Index Expanded (SCIE), PubMed and Scopus databases, and covering the period 1972 to 2014. The authors used a combination of tools including Ravar Matrix, NetDraw to map coauthorship networks and VOSViewer, a software to draw co-word maps. Using basic network properties such as closeness, betweenness, eigenvector centrality measures, they identified structural holes, active authors, analyzed the structural indices of their network and evaluated the trend of published articles. One important limitation of their study was the attempt to manually standardize Iranian authors' names and the keyword based search leading to the lack of

comprehensiveness of the search results.

A similar study of Iranian researchers on Medical Parasitology using NetDraw and the UCINET software package was also reported by Sadoughi et al. [71]. The study used basic network analysis to identify prolific researchers in the field of Medical parasitology by collecting 1048 published documents of all types in the field from 1972 to 2013 from the Web Of Science. The study identified aspects of scientific collaborations to help policy makers in the medical parasitology research area. A Brazilian study reported in the literature used the same methodological approach to generate new tools to help the Brazilian research fund to better select and prioritize research proposals [25]. Publication records were collected from the Web Of Knowledge (also known as Web Of Science) scientific database on seven neglected tropical diseases. Co-authorship networks were generated for each disease and analyzed using Pajek and NetDraw, a tool of the UCINET software package. The text-mining was implemented using the VantagePoint software. The results generated new information leading to better design and strategic planning and implementation of a research funding program. This study further supports that traditional criteria to fund research such as research productivity or impact factor of scientific journals are not valuable indicators for grant selection in low productivity neglected tropical diseases research areas. This Brazilian study is one of the few that focused on co-authorship network in the fields of neglected tropical diseases and the vast field of tropical infectious disease.

In an attempt to promote cooperative and translational research initiatives, another study investigated the state of scientific collaboration on Chagas disease research [26]. The

study presented the analyzis of the scientific literature on Chagas disease published in the PubMed database between 1940 and 2009. On a total of 13,989 documents retrieved, the authors applied bibliometrics, social network analysis, and clustering methods implemented via the Pajek program to analyze the evaluation of collaboration patterns and to identify influential research groups. The results revealed a dramatic increase in research collaborations. As in Newman [21], this study concluded that the co-authorship network of Chagas disease constitutes a "small world" network characterized by a high degree of clustering. Another important remark is the scarcity of African co-authorship network studies. Our review only identified the study by Toivanen and Ponomariov [67] who focused on research collaboration patterns in the African regional systems with less insights into specific research areas.

The majority of the studies reviewed above implemented their analyses using the Pajek program [72] or the UCINET software package which has the built-in NetDraw tool [73]. The Pajek program is suitable for the analysis and visualization of large networks. It has Graphical user interface and has other features including multidiemnsional scaling and structural analysis. Unlike the Pajek program, the UCINET software package has built-in advanced features and can handle networks which size up to 10,000 nodes, and accepts a large number of network file format including the pajek format. In their entirety, the studies reviewed above applied descriptive, basic social analysis methods.

Recently, Zhang [74] proposed a complex approach to social network analysis, emphasizing only on link prediction, one of the network topology inference questions. Her approach involved the development of a computationally efficient solution based on machine learning

techniques such as naive bayes, support vector machine, K-nearest neighbor implemented in the data mining software Weka [75] and the Python package Scikit [76]. The approach was tested on different datasets including a citation network, a co-authorship network and a protein-protein network. Quite often, these methods are not perfect since they failed to correctly tease out unreliable nodes from reliable ones, compromising the reliability of the network. However, new methodological approaches to scientific co-authorship network analysis are emerging to address those limitations. For example, Oliveira et al. [77] proposed a Bayesian approach to the analysis of such networks. Yet another limitation worth noting is that none of the studies reviewed above applied dynamic network analyses such as dynamic time series analysis or longitudinal network analysis [61].

2.3 Visualization tools for Co-authorship Networks

Various authors have proposed diverse tools for specifically visualizing and exploring co-authorship network data. One of such tools has been reported by Liu and colleagues [78] who proposed an author navigator application for visual examination of co-authorhip networks. In their conception of the toolkits, the authors combined a web based application tool for the interactive navigation of the network and a Java based backend swing application for the management of CGI requests. To support Brazilian researchers, Barbosa and colleagues proposed **VRRC**, a web based tool for the visualization and recommendation of co-authorship network [79]. According to its developers, **VRRC** provides an interactive visualization, an overview of the collaborations over time, and recommendations

to initiate new collaborations and reinforce existing ones. **VICI**, another co-authorship visualization tool was proposed by Odoni and colleagues [80]. **VICI** combined a Python based backend system for the extraction and management of the network data and a web based frontend using Flask [81] to display the network. The visualization of the network was finally rendered using the Javascript D3.js [82] library. **NeL**^{2}, a general purpose tool for the visualization of networks as a layered network diagram was proposed by Nakazono, Misue, and Tanaka [83]. They applied their tool to the visualization of co-authorship networks to visualize transitions in the network over a period of time, as well as various co-authorship data.

Another framework, the WebRelievo system was proposed for the visualization of the evolutionary processes of Web pages [84]. Other techniques were also proposed for the visualization of co-citation networks [85], and for the visualization of the relationship of scientific literature [86]. In addition to their inability to display large networks, those proposed tools are limited by their lack of interactivity and their inability for the end user to easily control the display. We therefore could not just re-use any one of them in this dissertation.

Chapter 3

A review of past approaches to the analysis of bibliometric data and co-authorship networks

To create and analyze graphs representing co-authorship of research publications this dissertation will rely on two types of methods. The first are methods for recognition when two superficially different representations of an author's name correspond to the same author, which is called "author name disambiguation". The second set of methods relate to the representation of graphs and different measures used to quantify the importance of relationships among the components of the graph (including vertices and subgraphs). In the following sections we will overview both types of work.

3.1 Author Name Disambiguation

Author Name Disambiguation (AND) is required because multiple names can refer to the same author, many authors may share the same name due to abbreviations, name misspellings, or identical names in publications [87]. AND remains an important research focus in the computer science community, prompting to proposed solutions to control authorship with manual curation via participative individual and community effort such as the Author-ity project [88], DBLife [89], the Open Researcher & Contributor ID (ORCID) [90], authorclaim.org, or researcherID.com. While most co-authorship analysis studies have tended to use a manual curation of AND [91], automatic approaches to AND involving supervised and unsupervised machine learning methods have also been proposed [92, 93]. Unfortunately, the proposed solutions presented above are still in their infancy. They have several limitations in that they often target a unique bibliographic database and do not usually contain old or relatively recent records. The Author-ity database for example in its last release (as of June 2018), only includes PubMed and Medline AND records up to September 2008. Here, because our data span from 1996 to 2016, we leveraged on the work of Bilenko [94] using an automatic, supervised fuzzy matching machine learning approach to disambiguate and normalize the bibliographic information collected (See section 4.2.1).

3.2 Basic Descriptive Analysis for Network Graphs

Vertex and edge characteristics are fundamental elements of network characterization. These characteristics are centered upon vertex and edge centrality measures. Although a vast number of different centrality measures have been proposed for the descriptive analysis of network graphs, the most common vertex and edge centrality measures are:

- Degree centrality: It is defined as the number of ties to a given author.
- Betweenness centrality: it is the number of shortest paths between other pairs of vertices that go through a particular vertex. It relates to the perspective that importance relates to where a vertex is located with respect to the paths in the network graph. According to Freeman [95], it is defined as:

$$c_B(v) = \frac{\sigma(s,t|v)}{\sum_{s \neq t \neq v \in V} \sigma(s,t)} \tag{3.1}$$

  where $\sigma(s,t|v)$ is the total number of shortest paths between vertices s and t that pass through vertex v, and $\sigma(s,t)$ is the total number of shortest paths between s and t regardless of whether or not they pass through v.

- Closeness centrality: the number of steps required for a particular author to access every other authors in the network. It captures the notion that a vertex is central if it is close to many other vertices. Considering a network $G = (V, E)$ where V is the set of vertices and E, the set of edges, the closeness centrality $c_{Cl}(v)$ of a vertex

v is defined as:

$$c_{Cl}(v) = \frac{1}{\sum_{u \in V} dist(v,u)} \tag{3.2}$$

where $dist(v,u)$ is defined as the geodesic distance between the vertices $u, v \in V$.

- Eigenvector centrality: degree to which an author is connected to other well connected authors in the network. It seeks to capture the idea that the more central the neighbors of a vertex are, the more central that vertex itself is. According to Bonacich [96] and Katz [97], the Eigenvector centrality measure is defined as:

$$c_{E_i}(v) = \alpha \sum_{\{u,v\} \in E} c_{E_i}(u) \tag{3.3}$$

  Where the vector $\mathbf{c}_{E_i} = (c_{E_i}(1), \ldots, c_{E_i}(N_v))^T$ is the solution to the eigenvalue problem $\mathbf{A}\mathbf{c}_{E_i} = \alpha^{-1}\mathbf{c}_{E_i}$, where $\mathbf{A}$ is the adjacency matrix for the network G. According to Bonacich [96], an optimal choice of α^{-1} is the largest eigenvalue of $\mathbf{A}$

- Brokerage: degree to which a vertex occupy an advantageous position such that they can broker interactions between other vertices in the network.

- Edge betweenness centrality extends from the notion of vertex centrality. It reflects the number of shortest paths traversing that edge.

3.2.1 Characterizing Network cohesion

There are many techniques to determine network cohesion [98]:

- Cliques: According to Kolaczyk and Csárdi [98], cliques are defined as complete subgraphs such that all vertices within the subset are connected by edges.

- Density: Defined as the frequency of realized edges relative to potential edges, the density of a subgraph H in G provides a measure of how close H is to be a clique in G. Density values vary between 0 and 1:

$$den(H) = \frac{|E_H|}{|V_H|(V_H - 1)/2} \tag{3.4}$$

- Transitivity: The transitivity of G is a measure of the relative frequency of G defined as:

$$cl_T = \frac{3\tau_\Delta(G)}{\tau_3(G)} \tag{3.5}$$

  where $\tau_\Delta(G)$ is the number of triangles in G, and $\tau_3(G)$ is the number of connected triples (sometimes referred to as 2-star). This measure is also referred to as the fraction of transitive triples. It represents a measure of global clustering of G summarizing the relative frequency with which connected triples close to form triangles [98].

- Connectivity, Cuts, and Flows: The concepts of vertex and edge cuts is derived from the concept of vertex (edge) connectivity. The vertex (edge) connectivity of a

graph G is the largest integer such that G is k-vertex- (edge-) connected [98]. These measures helped assess the most important vertices for information flow and the long-term sustainability of each network. Since co-authorship networks are undirected graphs, the concept of weak and strong connectivity is irrelevant. A graph G is said to be connected if every vertex in G is reachable from every other vertex. Usually, one of the connected components can dominate the others, hence the concept of giant component. The giant component characterizes the connectedness of the vertices in the network.

- Graph Partitioning: Regularly framed as a community detection problem, graph partitioning identifies cohesive subsets of vertices generally well connected among themselves and well separated from the other vertices in the network graph. Two established methods of graph partitioning are Hierarchical clustering (agglomerative vs divisive) and Spectral clustering [98].

3.3 Modeling of Network data

The purposes of network graph modeling are to test significance of the characteristics of observed network graphs, and to study proposed mechanisms of real-world networks such as degree distributions and small-world effects [98]. A model for a network graph is a collection of possible graphs $\mathscr{G}$ with a probability distribution $\mathbb{P}_\theta$ defined as:

$$\{\mathbb{P}_\theta\ (G), G \in \mathscr{G} : \theta \in \Theta\} \tag{3.6}$$

where θ is a vector of parameters ranging over values in Θ.

Given an observed network graph G^{obs} and some structural characteristics $\eta(\cdot)$, our goal is to assess if $\eta(G^{obs})$ is unusual. We then compare $\eta(G^{obs})$ to collection of values $\{\eta(G) : G \in \mathscr{G}\}$. If $\eta(G^{obs})$ is too extreme with respect to this collection, then we have enough evidence to assert that $\eta(G^{obs})$ is not a uniform draw from $\mathscr{G}$.

3.3.1 Mathematical models for Network Graphs

There are mainly four proposed mathematical models for network graphs:

- Classical Random Graph Models: First established by Erdős and Rényi [99–101], it specifies a collection of graphs $\mathscr{G}$ with a uniform probability $\mathbb{P}(\cdot)$ over $\mathscr{G}$. A variant of this model called the Bernoulli Random Graph Model was also defined by Gilbert [102].

- Generalized Random Graph Models: These models emanated from the generalization of Erdős and Rényi's formulation, defining a collection of graphs $\mathscr{G}$ with prespecified degree sequence.

- Mechanistic Network Graph Models: These models mimic real-world phenomena and include Small-World Models commonly referred to as "six-degree separation". It was introduced by Watts and Strogatz [103] and have since received a lot of interests in the existing literature especially in Neuroscience. Small-world networks usually exhibit high levels of clustering and small distances between vertices. Classical models are not fit to better represent such behaviors since they usually display

low levels of clustering and small distance between vertices. Examples of known small-world networks include the network of connected proteins or the transcriptional networks of genes [104]. A variant of Small-World models is the Preferential Attachment Models defined based on the popular principle of "the rich get richer". Preferential attachment models gained fascination after the work of Barabási and Albert who studied the growth of the World Wide Web [105]. Examples of Preferential Attachment networks include that of the World Wide Web and the scientific citation network [106, 107]. An important characteristic of these models is that as time tend to infinity, there degree distribution tends to follow a power law.

3.3.2 Statistical models for Network Graphs

Although mathematical models tend to be simpler than statistical models, the latter allow model fitting and assessment. Various classes of network graph models have been proposed. Here, we present the three main classes of statistical network models and a version of ERGM adapted to temporal snapshots:

3.3.2.1 Stochastic Block Model

Blockmodel is a statistical method to identify, in a given network, clusters or classes of authors that share structural characteristics [108, 109]. Each such cluster forms a position. The units within a cluster have the same or similar connection patterns. Given a graph $G = (V, E)$ and its adjacency matrix $\mathbf{Y}$, for two distinct nodes $i, j \in V$, the block model

defined by Kolaczyk and Csárdi [98], specifies that each element Y_{ij} of $\mathbf{Y}$ is conditional on the class label q and r of the vertices i and j. The model has the form:

$$Pr(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa}\right) exp\left\{\sum_{q,r} \theta_{qr} L_{qr}(\mathbf{y})\right\} \tag{3.7}$$

where L_{qr} is the number of edges in the observed graph $\mathbf{y}$ connecting vertices of classes q and r, θ_{qr} is the parameter estimates, and κ is a normalization constant defined as:

$$\kappa = \sum_{\mathbf{y}} exp\left\{\sum_{q,r} \theta_{qr} L_{qr}(\mathbf{y})\right\} \tag{3.8}$$

Stochastic block model (SBM) originated from the ideas that equivalent units can be grouped together. There are three definitions of equivalences which are structural, automorphic and regular [61]. In practice, the differences in types of equivalence tend to blur when stochastic block modeling is applied to real networks.

3.3.2.2 Exponential Random Graph Model

Also referred to as p* models, Exponential Random Graph Models (ERGMs) are probability models for network designed in analogy to Generalized Linear Models (GLMs) [98]. ERGMs have gained increasing interests especially in modeling social networks. Robins et al. [110] provide a nice introduction to ERGM as well as a general framework for ERGM creation which we closely followed in this dissertation.

Given a random graph $G = (V, E)$, for two distinct nodes $i, j \in V$, we define a random

binary variable Y_{ij} such that $Y_{ij} = 1$ if there is an edge $e \in E$ between i and j, and $Y_{ij} = 0$ otherwise. Since co-authorship networks are by definition undirected networks, $Y_{ij} = Y_{ji}$ and the matrix $\mathbf{Y} = [Y_{ij}]$ represents the random adjacency matrix for G. The general formulation of ERGM is therefore:

$$Pr(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa}\right) exp\left\{\sum_{H} \theta_H g_H(\mathbf{y})\right\} \tag{3.9}$$

where each H is a configuration, a set of possible edges among a subset of the vertices in G and $g_H(\mathbf{y}) = \prod_{y_{ij} \in H} y_{ij}$ is the network statistic corresponding to the configuration H; $g_H(\mathbf{y}) = 1$ if the configuration is observed in the network $\mathbf{y}$, and is 0 otherwise. θ_H is the parameter corresponding to the configuration H (and is non-zero only if all pairs of variables in H are assumed to be conditionally dependent); κ is a normalization constant defined as:

$$\kappa = \sum_{\mathbf{y}} exp\left\{\sum_{H} \theta_H g_H(\mathbf{y})\right\} \tag{3.10}$$

3.3.2.3 Temporal Exponential Random Graph Model

The Temporal Exponential Random Graph Model (TERGM) is an extension of the ERGM described in section 4.4.2.2 proposed by Hanneke, Fu, and Xing [111] from the work of Robins and Pattison [112]. The TERGM was designed with the idea of accounting for inter-temporal dependence in longitudinally collected network data. For a full description of the TERGM, we refer the reader to Leifeld, Cranmer, and Desmarais [113].

3.3.2.4 Latent Network Model

Designed in analogy to Mixed Models, Latent Network Models (LNM) allow the incorporation of latent or unobserved variables in network modeling. These models specifically account for structural equivalence, to model hidden factors or information not available in the network. Kolaczyk and Csárdi [98] provide a formulation of LNM. Given the adjacency matrix $\mathbf{Y}$ of a graph $G = (V, E)$, for each element Y_{ij} of $\mathbf{Y}$, the latent variable model is of the form:

$$Y_{ij} = h(\theta, z_i, z_j, \epsilon_{ij}) \tag{3.11}$$

where θ is a constant, the ϵ_{ij} are independent and identically distributed pair-specific effects, and h is a symmetric function. The model assumes that each vertex $i \in V$ has a latent variable z_i. Considering observed covariates $\mathbf{Z}$, the probability of forming an edge between two nodes i and j $(i, j \in V)$ is independent of all other vertex pairs given values of latent variables, and is defined as:

$$Pr(\mathbf{Y}|\mathbf{Z}, \theta) = \prod_{i \neq j} Pr\left(Y_{ij}|z_i, z_j, \theta\right) \tag{3.12}$$

Chapter 4

Methodology

4.1 Overview

To attain objective 1, our methodological approach consists in performing descriptive analysis of the network data of each co-authorship network, following the methodology used by Newman et al. [21] and Ghafouri et al. [24]. For objective 2, we use clustering methods, and shortest path algorithms as explained by Newman [57, 58]. Next, we apply mathematical modeling to attain objective 3. Regarding objective 4, we apply advanced statistical modeling including dynamic or longitudinal network analysis methods as recommended by Mali et al. [61]. We use a number of visualization methods to display the results. Finally, we develop a prototype of co-authorship tool to predict future research collaboration ties using the best performing statistical models.

The methodology workflow is presented in figure 4.1.

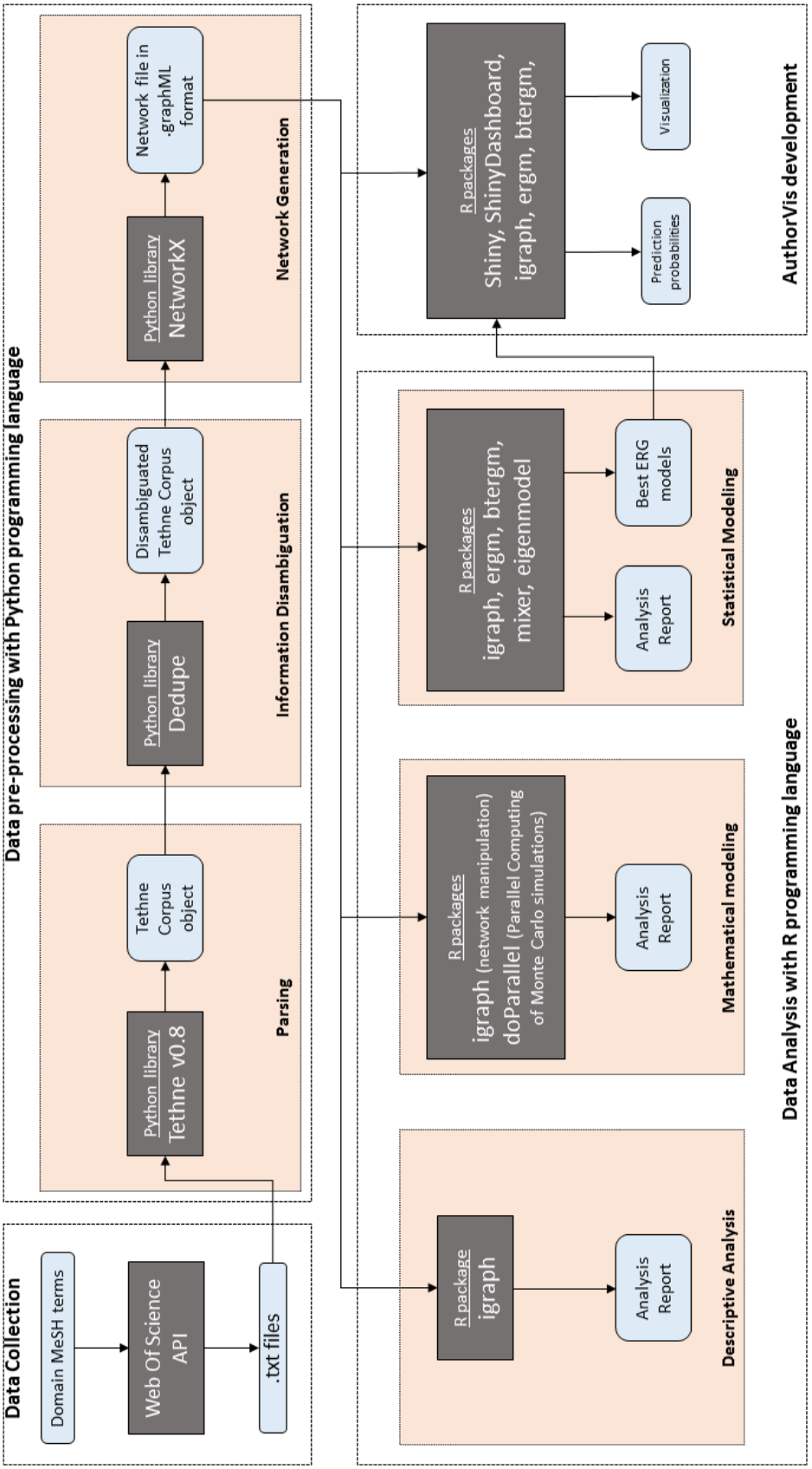


FIGURE 4.1: Methodology Workflow

4.2 Data Collection

Our research utilized secondary data collection techniques using the systematic literature search. We collected publication records indexed in the Thompson's Institute for Scientific Information Web Of Science (formerly known as the Web of Knowledge). For each disease domain, we searched the WOS databases using combinations of disease related MeSH terms. For the malaria research domain, we combined the following MeSH terms: "Malaria", "Anopheles", "Plasmodium" and "vector". The HIV/AIDS related MeSH terms are "HIV", "AIDS", "VIH", and "HIV infections". The TB related MeSH terms include "Tuberculosis", "Mycobacterium", and "Infection". We restricted the search to the period from 1996 to 2016 and to "Benin" for country. We manually screened the records in order to only select those published by Beninese authors, or papers published on each disease domain involving at least one author affiliated to a Beninese research institution. No restriction was placed upon the document types. For each disease domain, we first queried the WOS with each MeSH term independently, then combined the other terms so the query return the maximum number of results. The Full citations information containing the authors' names, their institutional affiliations, the year of publication, as well as the number of times the document was cited were recorded as bibliographic text files. After a second manual screening, only records that met the above listed inclusion criteria were finally selected. The selected records were saved in bibliographic text files and input to parser and functions for disambiguating the names of authors and other entities such as cities or research facilities.

4.2.1 Parsing and Information Disambiguation

We used Dedupe, a python library (obtained from `https://github.com/datamade/dedupe`) to disambiguate authors' names and assign a unique identification number to each author. We manually annotated 10% of the names and then trained the algorithm to automatically disambiguate the remaining of the entries. Dedupe is interactive and adjusts further annotations as the disambiguation process evolves. We evaluated our AND fuzzy matching machine learning method by computing Precision and recall metrics. Dedupe was also used to normalize and disambiguate other information such as research center affiliations, city, and country. At the end of the Information Disambiguation process, a disambiguated Tethne corpus object was generated and used as input to the co-authorship network generation processing.

4.2.2 Network Generation

Using NetworkX [114], another python library, we wrote a script taking the disambiguated Tethne corpus object as input to generate undirected multigraph co-authorship networks containing parallel edges. Each author or researcher from the disambiguated Tethne corpus represented a vertex. An edge was created between two authors (vertices) when they author a document together. Multiple parallel edges were created between two authors when they coauthor multiple papers together. Our script output NetworkX graph objects where vertices were defined by several attributes including name, affiliation, city, country, number of publication and total number of times cited. Edges had attributes

associated with them such as a unique identifier, the number of times a pair of authors was cited and the number of publications of a pair of authors. The undirected multigraph networkX objects were finally exported as .graphML files and used as input for data analyses.

4.3 Descriptive Data Analysis

For each co-authorship network, the numbers of authors, edges, and publications are plotted against the co-authorship years span. Using **igraph**, a network analysis package developed in R, each of the graphML files is converted into an igraph network object. For the descriptive analysis, we use the **igraph** package to compute the vertex degree and examined the degree distribution using both the natural frequency and the log scale degree distribution to characterize the type of distribution. We also computed vertex closeness, betweenness, eigenvector centrality measures and edge betweenness centrality measures to respectively identify the top 10 most connected authors, the top 10 broker authors, the top 10 network hubs, and the top 10 most important edges for information flow.

4.3.1 Characterizing Network cohesion

The extent to which subsets of authors are cohesive with respect to their relation in the co-authorship network was assessed through network cohesion. Specifically, we determined if

collaborators (co-authors) of a given author tend to collaborate as well, and what subset of collaborating authors tend to be more productive in the network. Using the **igraph** package, we conducted clique detection by computing the maximal cliques and their sizes, the density, and the transitivity. We also conducted a census of the connected components in each network, identify the giant component and characterize its size. Cut vertices were also computed to list the weak articulation points of each network.

The agglomerative hierarchical clustering method was used to identify clusters (or research communities) in the network. Finally, we generated a visualization of each co-authorship network weighting the vertex size by their betweenness values and assigning colors based on their cluster membership determined by the hierarchical clustering method.

4.4 Modeling of Network Data

4.4.1 Mathematical Modeling

We input the observed characteristics of each co-authorship network to an **igraph** function to perform 1,000 Monte-Carlo based simulations of the four different mathematical models for network graphs (Classical Random Graph, Generalized Random Graph, Watts-Strogatz Small-World, and the Barabási-Albert Preferential Attachment) presented in section 3.3.1. We assessed the significance of the observed characteristics by comparing them to those of the 1,000 simulated networks using a one sample Student's t-test. Characteristics we assessed significance for are the average shortest paths, the clustering coefficient and the number of communities detected by the hierarchical clustering methods.

4.4.2 Statistical Modeling

To model the complexity of the structure of each co-authorship network, we fit the SBM, the ERGM, the TERGM, and the LNM (presented in section 3.3.2) to each co-authorship network data. For each model, we computed and included in the model an important social network principle referred to as homophily which is defined in our network as the tendency of similar authors to collaborate. Another very important social network principle we also computed and included in the model, is the one of structural equivalence which is the similarity of network positions on the formation of collaboration ties in a

given network. We used the results from the static and temporal statistical network models listed above to verify the hypotheses in section 1.2. The purpose of this approach to network modeling is to unveil structural patterns driving collaboration tie formation in each co-authorship network.

4.4.2.1 Stochastic Block Model

We used SBM to both model each of the observed networks but also as a model based clustering technique. After fitting the SBM, we examined the posterior probability of class membership from the returned object. We then determined the class membership of each vertex class assignment based on the maximum a posteriori criterion. Class membership was added to the network as an additional nodal attribute. Subroutines of R package **mixer** [115–118] was used to fit and evaluate the SBM. **Mixer** used the Integration Classification Likelihood (ICL) criterion to select the number of classes fit to the observed network. We finally examined the summary plot generated by the **Mixer** package which contains the ICL, the degree distribution, the reorganized adjacency matrix, and the inter/intra class probability plots. While the ICL plot displays the optimal number of classes, the degree distribution helps assess the goodness-of-fit of the SBM to the observed data. The reorganized adjacency matrix plot shows the interactions between the classes of the network and the inter/intra class probabilities plot highlights the inter and intra interactions between the detected classes.

4.4.2.2 Exponential Random Graph Model

The R package **ergm** [119, 120] was used to fit ERGM to the observed networks. We used ERGM to model the network ties, the dependent variable as a function of nodal and dyadic attributes (covariates) such as the number of times an author was cited, the number of publications, the number of collaborators, the collaboration type as well as its community membership as determined by the SBM.

Given the high transitivity coefficient of this network, we also included transitivity as a network structural process. As recommended for ERGM model specification for undirected network, we investigated homophily which is the tendency of similar author to collaborate. We also included factor attribute effect in the model.

Several models containing nodal, dyadic and structural terms were fit to the observed network data. The first model we fit is a naive model containing only the ERGM "edge" term. This model is nothing but the Bernoulli random graph model [99]. We then fit another model containing only nodal and/or dyadic terms. Third, we fit a structural model containing only high-order terms representing network statistics such as triangles, k-stars, geometrically weighted edge-wise shared partner distribution and many more [98, 110].

Model log-likelihood, the Akaike's Information (AIC) and the Bayesian Information (BIC) criteria were used to select the best ERGM. The best model was selected based on the lowest AIC, or the lowest BIC, and the highest log-likelihood. Usually, AIC and BIC decrease or increase together. In case of conflicting trend in AIC and BIC values, the log-likelihood was used to select the best model. We checked for model diagnostics by computing and inspecting the Goodness-Of-Fit visualization for the best model using a

subroutine of the **ergm** package. A maximum of 1,000 iterations and 1,000 simulations was set as parameters to the ERGMs.

4.4.2.3 Temporal Exponential Random Graph Model

All Temporal Exponential Random Graph Models (TERGMs) were fit using the Markov Chain Monte Carlo Maximum Likelihood Estimation (MCMC-MLE) implemented in the **btergm** R package [113]. We divided each network in different snapshots spanning different intervals of time using a manual process such that the temporal snapshots are not overly dense or sparse early on or in later time periods. We used **igraph** to visualize and manually verified that the temporal snapshots are balanced across the time periods. We then modeled the network ties, the dependent variable as a function of nodal and dyadic variables. Dyadic stability and delay reciprocity memory TERGM terms were also included in the model. To check whether there is a linear trend in collaboration tie formation, we also included a linear time covariate in the model. We accounted for network structural predictors and homophily on the type of collaboration. Model log-likelihood, the Akaike's Information (AIC) and the Bayesian Information (BIC) criteria were used to select the best TERGM corresponding to the lowest AIC or BIC, and highest log-likelihood. AIC and BIC are estimates of a function of the posterior probability of a model being true. Under a Bayesian setup, a lower BIC or AIC means that a model is more likely to be the true model.

To evaluate the extent to which the final model captures the endogenous properties and processes of the observed network, we assessed model diagnostics, by inspecting

the within-sample and out-of-sample goodness-of-fit visualization computed from a subroutine of the **btergm** package. For the out-of-sample goodness-of-fit, we estimated the model on the first network snapshots leaving out the last network snapshot in the series. We simulated $1,000$ networks from the model and assessed how the simulated networks predicted the left out network. As described by Desmarais and Cranmer [121], we also provided a micro-interpretation of the final TERGM.

4.4.2.4 Latent Network Model

Hoff [122] suggested an approach based upon the principles of eigen-analysis of specifying latent variables which we followed in this dissertation. The R package **eigenmodel** developed by Hoff [123] was used to fit the LNM to the observed networks. We fit LNM with both no pair-specific and pair-specific covariates such as the type of collaboration and community assignment from the SBM. The rationale of fitting the pair-specific models with those two variables is supported by our third hypothesis which states that collaboration ties in each co-authorship network are driven by homophily in terms of community membership and/or collaboration type. We also fit other pair-specific covariates model using nodal and dyadic covariates. We visualized and compared the co-authorship network using 3 dimensional layouts determined according to the inferred latent eigenvectors in each model. Finally, we used a 5-fold cross-validation method to assess the goodness-of-fit of each model which we compared using ROC curves via the R package **ROCR** [124].

The creation of the co-authorship tool is presented in Chapter 8.

Chapter 5

Results: The Malaria Co-authorship Network

5.1 Data

The search was conducted using combinations of Malaria related MeSH terms including ”malaria”, ”Anopheles”, ”Plasmodium” and ”vector”. The final query set (Table 5.1) returned 685 records. After screening, 424 documents met the selection criteria. On average, there was 10.67 authors per published document.

After the Author Name Disambiguation, we identified 1792 unique authors with a precision of 99.87% and a recall of 95.46%. The generated multigraph co-authorship network therefore contained 1792 vertices (authors) and 116,388 parallel edges (collaborations).

TABLE 5.1. Malaria Bibliographic Search Queries.

Set	Queries	Results
#1	TOPIC: (malaria) OR TOPIC: (mosquito), Refined by: COUNTRIES/TERRITORIES: (BENIN)	513
#2	TOPIC: (malaria) OR TOPIC: (mosquito) OR TOPIC: (anopheles), Refined by: COUNTRIES/TERRITORIES: (BENIN)	529
#3	TOPIC: (malaria) OR TOPIC: (mosquito) OR TOPIC: (anopheles) OR TOPIC: (plasmodium) OR TOPIC: (bednet), Refined by: COUNTRIES/TERRITORIES: (BENIN)	544
#4	TOPIC: (malaria) OR TOPIC: (mosquito) OR TOPIC: (anopheles) OR TOPIC: (plasmodium) OR TOPIC: (net) OR TOPIC: (vector), Refined by: COUNTRIES/TERRITORIES: (BENIN)	685
Final Set	#1 OR #2 OR #3 OR #4	685

The evolution of the published Malaria related documents, authors and collaborations from January 1996 to December 2016 is presented on figure 5.1.

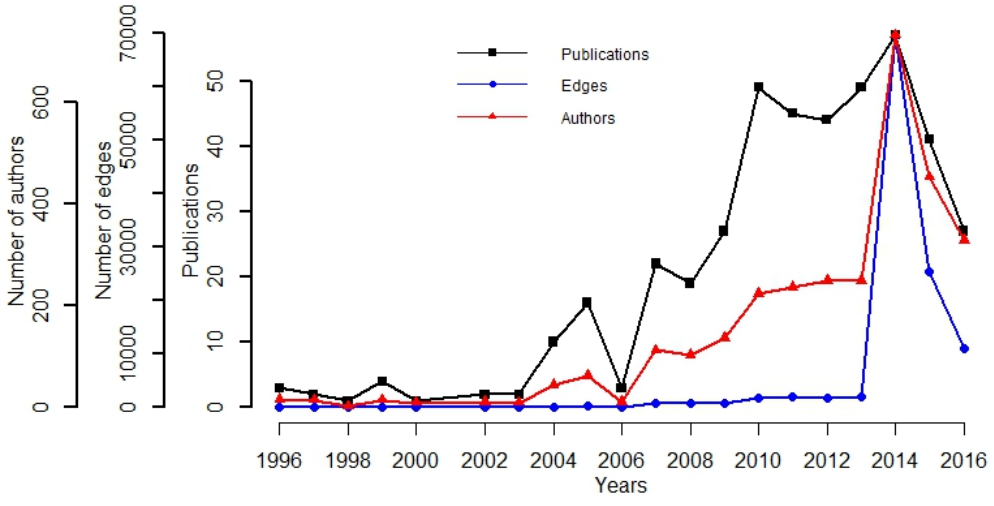


FIGURE 5.1: Evolution of the published Malaria related documents, authors and collaborations from January 1996 to December 2016

5.2 Descriptive Data Analysis

The degrees of the multigraph network range between 1 and 1338 with an average degree distribution of 106.46. We noted in addition, a substantial number of vertices with low degrees (Fig. 5.2). There was also a non-trivial number of vertices with higher order of degree magnitudes. A log scale distribution of the degrees demonstrate that the vertex degrees tend to follow a heavy-tail distribution.

After we convert the multigraph network in a weighted graph, it results in a simple graph of 1792 vertices and 95,787 weighted edges. Mean Closeness centrality ranges between 3.118×10^{-7} and 5.152×10^{-6} with a median of 5.112×10^{-6}. This measure suggests a highly right-skewed distribution. Betweenness measures range between 0 and 245600 with a median of 1985. A network visualization with the vertices' size proportional to betweenness centrality measures clearly reveals the presence of broker authors (Table 5.2). The median eigenvectors measure is 0.005, its mean is estimated at 0.09. Eigenvectors measures reveal the presence of multiple cluttered authors suggesting the presence of closed collaboration groups. Table 5.2 presents a list of the 10 authors with the highest Eigenvectors values.

The computation of edge betweenness identifies co-authorship collaborations that are important for the flow of information. In Table 5.2, We present the top 10 most important collaborations for the flow of information in the Malaria co-authorship network in Benin.

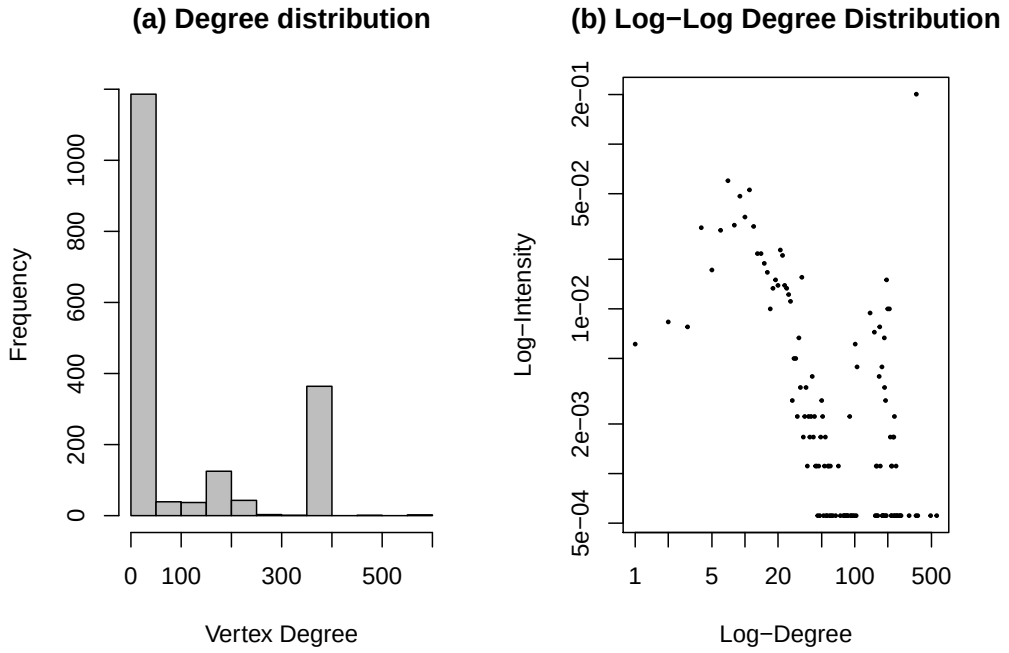


FIGURE 5.2: Degree distribution of the Malaria co-authorship network

5.2.1 Network Cohesion

A total of 365 maximal cliques are identified in the network among which 9 cliques of size 2, 14 cliques of size 3, 155 cliques of size 8, and 142 cliques of size 7. Larger maximal cliques sizes range from 102 authors to 365 authors and are all found once across the network.

The malaria co-authorship network has a density of 0.0596 and a transitivity of 0.965 indicating that 96.5% of the connected triples in the network are close to form triangles. The transitivity metrics is a measure of the global clustering of the network.

The network is not connected and a census of all the connected components within the network reveals the existence of a giant component that dominates all the other connected components. This giant component includes 94% (1686 vertices) of all the vertices in the

TABLE 5.2. List of the most important authors and collaborations in the Malaria co-authorship network

Top 10 Brokers
MASSOUGBODJI ACHILLE
HAY SIMON I
KAREMA CORINE
SANNI AMBALIOU
KENGNE ANDRE PASCAL
AKOGBETO MARTIN
NDAM NICAISE TUIKUE
MALIK ELFATIH M
DABIRE K ROCH
DELORON PHILIPPE
Top 10 most connected authors (Top 10 network hubs)
MASSOUGBODJI ACHILLE
KAREMA CORINE
GONZALEZ RAQUEL
MENENDEZ CLARA
DALESSANDRO UMBERTO
OGUTU BERNHARDS R
FAUCHER JEANFRANCOIS
BASSAT QUIQUE
MARTENSSON ANDREAS
HAY SIMON I
Top 10 most important edges for information flow
DABIRE K ROCH – KENGNE ANDRE PASCAL
BALDET THIERRY – KENGNE ANDRE PASCAL
AKOGBETO MARTIN – MALIK ELFATIH M
AVLESSI FELICIEN – MOUDACHIROU MANSOUROU
AKOGBETO MARTIN – AVLESSI FELICIEN
MASSOUGBODJI ACHILLE – RAHIMY MOHAMED CHERIF
DIABATE ABDOULAYE – KENGNE ANDRE PASCAL
GARCIA ANDRE – SANNI AMBALIOU
KAREMA CORINE – MALIK ELFATIH M
HAY SIMON I – MALIK ELFATIH M
Weak articulation points
NOEL VALERIE
DJOGBENOU LUC
ZOHOUN I
SANNI AMBALIOU
EDORH ALEODJRODO PATRICK
ALLABI AUREL
HOUNKONNOU MAHOUTON NORBERT
FAYOMI BENJAMIN
KINDEGAZARD DOROTHEE A
DJOUAKA ROUSSEAU
RAHIMY MOHAMED CHERIF
BALDET THIERRY
DOSSOUGBETE L
GARCIA ANDRE
MASSOUGBODJI ACHILLE
AKOGBETO MARTIN

network with none of the other components alone carrying less than 1% of the vertices in the network (Fig. 5.3).

The assessment of information flow in the network via cut vertices reveal the existence of 16 authors as the most vulnerable vertices in the network. Table 5.2 lists the authors

that constitute the weak articulation points in the malaria co-authorship network. Cut vertices are crucial to the sustainability of networks [98].

The agglomerative hierarchical clustering method identifies 23 research communities (or clusters) in the network. Sizes of the clusters range between 2 and 570 with large research communities containing between 202 and 569 authors. Medium size research communities contain between 10 and 62 authors. Only 7 out of the 23 research communities identified are part of the giant component. Figure 5.3 displays the giant component of the network with each different colors representing each of the 7 research communities.

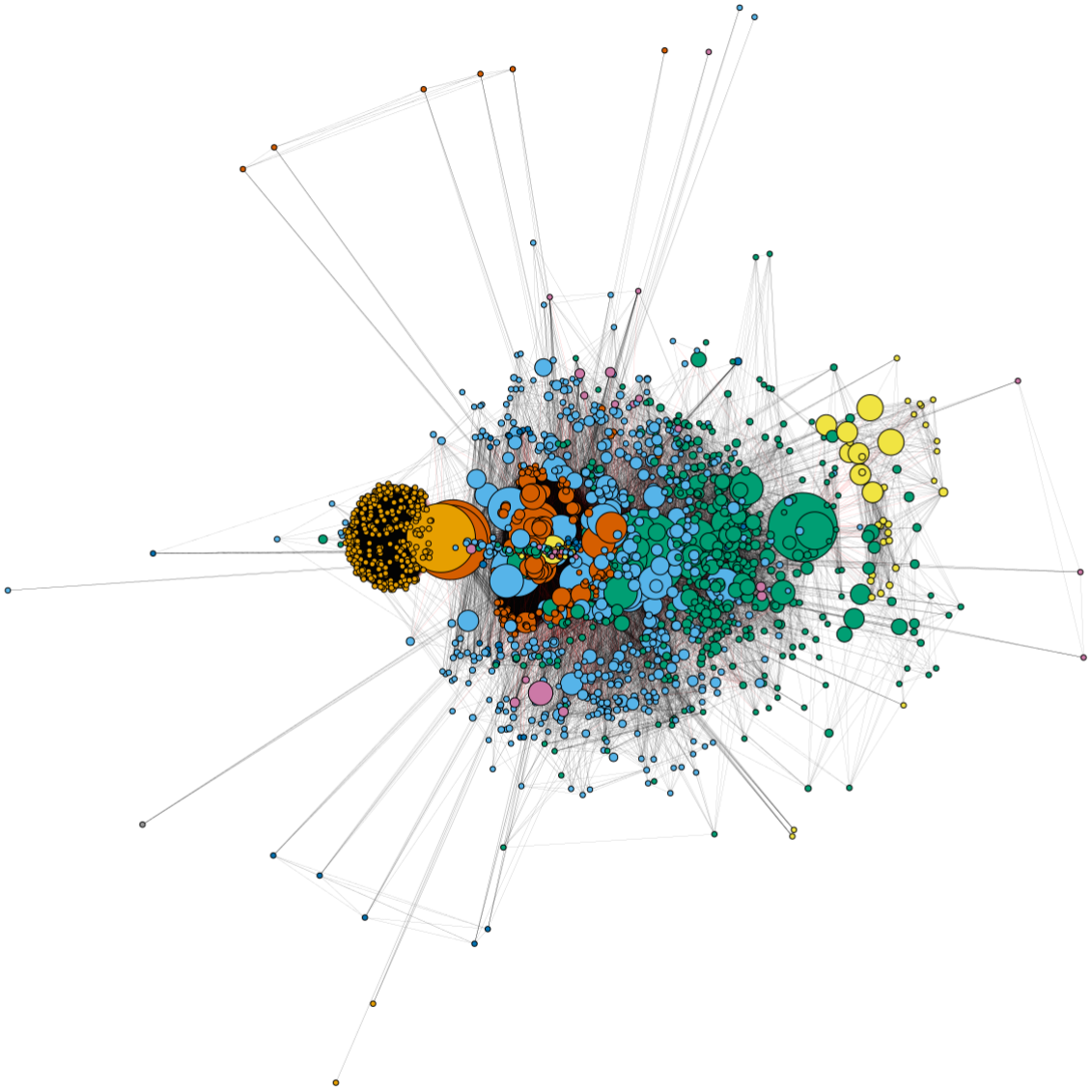


FIGURE 5.3: Malaria co-authorship network – Main component.
Authors (vertices) of the same color belong to the same research community or cluster

5.3 Modeling

5.3.1 Mathematical Modeling

The hierarchical clustering method of community detection algorithm has identified 23 different clusters/communities in the co-authorship network out of which 7 form a giant component. One of the question of interest in this section is whether the number of communities detected is expected or not. The results of 1,000 Monte Carlo based simulations to test the significance of this observed characteristic are presented on figures 5.4 and 5.5. Figure 5.4 clearly demonstrates that the number of communities detected is unusual from the perspective of both Classical random graphs and generalized random graphs (p-value < 0.0001). From the Classical random graph model, the expected number of communities is 3.934 (95%CI: 3.90 – 3.97). Similarly, the expected number of communities from the generalized random graph model is 7.501 (95%CI: 7.39 – 7.61).

Figure 5.5 displays the number of detected research communities using the Barabási-Albert's preferential attachment and the Watts-Strogatz models. Supprisingly enough, the observed number of communities is also extreme per both models (p-value < 0.0001). The expected number from the Watts-Strogatz model simulations is 3.056 (95%CI: 3.04 – 3.07) and 45.569 (95%CI: 45.42 – 45.72) from the Barabási-Albert model simulations.

We also compared the clustering coefficient and the average shortest-path length. The observed clustering coefficient is 0.9645. Surprisingly, there is substantially more clustering in our malaria co-authorship network than expected from all 4 mathematical models

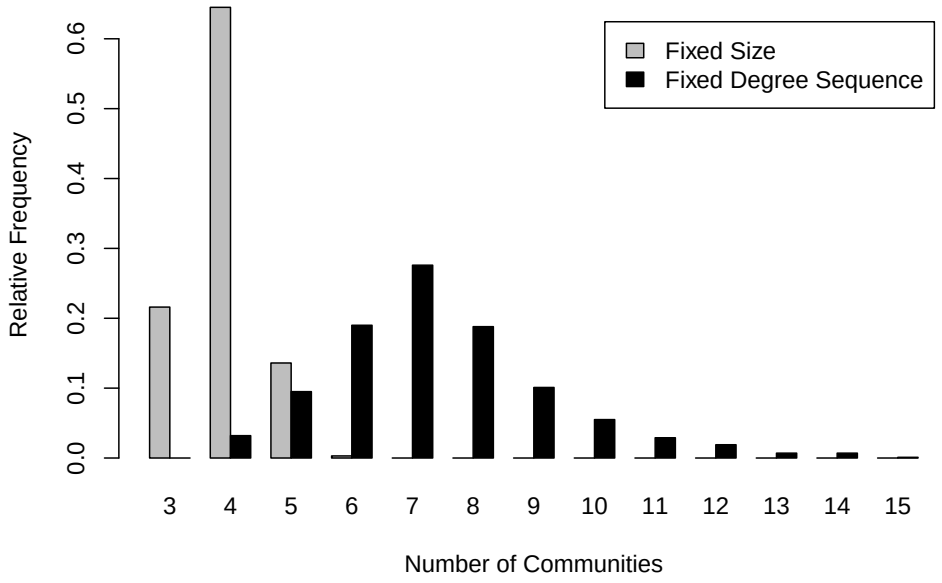


FIGURE 5.4: Monte-Carlo simulations: Number of detected communities by the random graph models

(p-value < 0.0001). The expected clustering coefficient is 0.0596 (95%CI: 0.05963068 – 0.05964648) and 0.4334 (95%CI: 0.4333912 – 0.4334522) respectively for the classic random graph and the generalized random graph models.

Similarly, The Watts-Strogatz Small World model expected clustering is 0.7464 (95%CI: 0.7464326 – 0.7464356).

We observed an average shortest-path length of 2.99 in the malaria co-authorship network. This observed shortest-path length is significantly larger than what is expected from the random graph models (p-value < 0.0001) and significantly lower than what is expected from Watts-Strogatz small world model and the Barabási-Albert preferential attachment model (p-value < 0.0001).

The average shortest-path length is 1.94 (95%CI: 1.941955 – 1.941960) and 2.26 (95%CI:

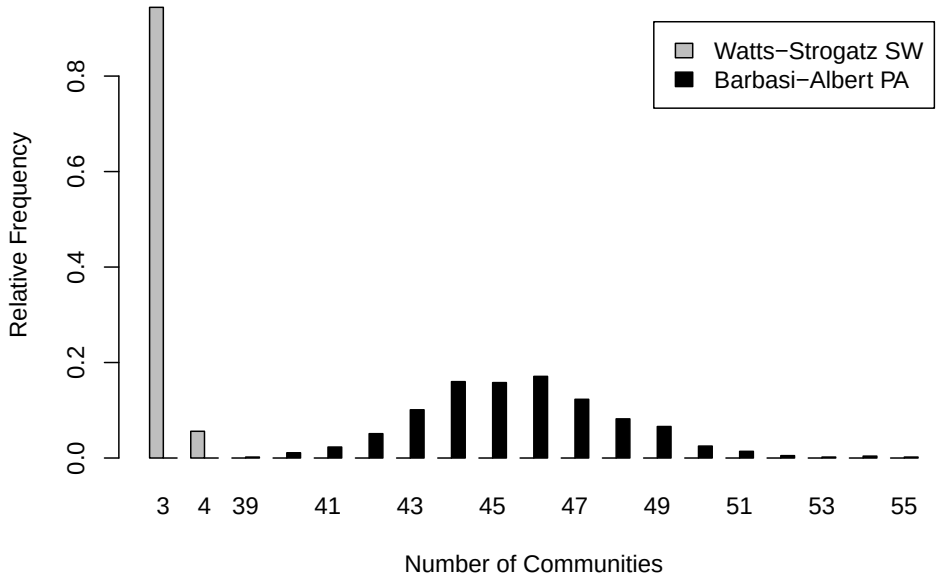


FIGURE 5.5: Monte-Carlo simulations: Number of detected communities by the Watts-Strogatz and the Barabási-Albert models

2.259468 2.259586) respectively for the classic random graph and the generalized random graph models.

For the Watts-Strogatz small world and the Barabási-Albert models, the average shortest-path length is respectively 3.83 (95%CI: 3.81 – 3.86) and 9.17 (95%CI: 9.14 – 9.21).

All simulations were also performed on the giant component of the network and led to similar outcomes.

5.3.2 Statistical Modeling

5.3.2.1 Stochastic Block Model

The ICL plot on figure 5.6 shows that the malaria co-authorship network has been fit with 39 classes by the SBM with a degree of latitude of 30 to 39 classes being reasonable. The degree distribution of the fitted SBM (blue curve) provides a decent description of the observed distribution (yellow histogram). In the inter/intra class probabilities network, the vertices correspond to the 39 classes detected by the SBM. The vertex sizes are proportional to the number of authors assigned to each class. Each vertex is further broken down in a pie chart with each portion reflecting the relative proportion of the types of collaboration. Yellow represents the proportion of authors of international affiliations, orange represents regional authors who are affiliated with African institutions other than Beninese institutions, and green for authors affiliated to Beninese research institutions. In general, we observe a dominance of international and regional researchers over national researchers across all detected clusters.

A close look at the reorganized adjacency matrix, reveals the presence of 4 larger classes (classes number 2, 4, 10 and 27) and 35 other classes of smaller sizes. One of the larger class (class 27) displays a tendency of its members to only establish collaboration ties between themselves. This class seems to have the characteristics of a clique. Examination of the distribution of each class by their type of collaboration (Figure 5.7) indicates that this class of authors (class 27) is primarily made of international contributors to the malaria research effort in Benin. Although members of this class seem to have rare

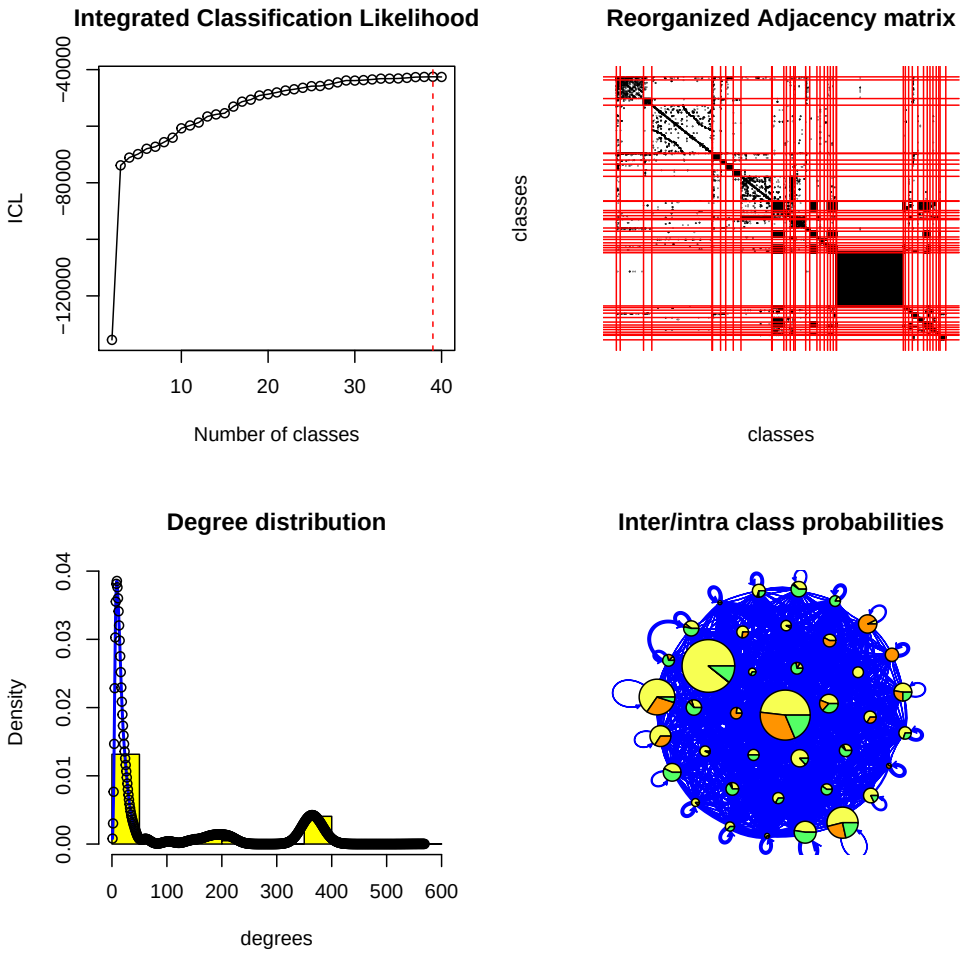


FIGURE 5.6: Summary of the goodness-of-fit of the SBM analysis on the Malaria co-authorship network.

collaboration ties with members of other classes, we also notice the presence of very few broker authors as national liaisons between this class 27 and another larger class (class 2). Though, it also appears in the other three larger classes that the authors tend to primarily collaborate within their respective classes, they also tend to collaborate with authors of other classes.

Figure 5.7 also shows that the co-authorship malaria network in Benin is dominated by

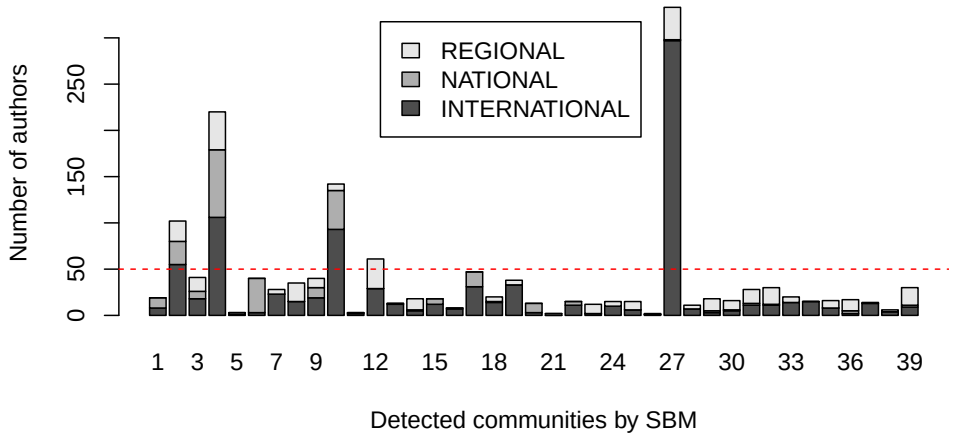


FIGURE 5.7: Distribution of national, international and regional authors by communities detected by the SBM in the Malaria network.

international researchers with national contributors unevenly distributed across the detected research communities. In order to better explain the inter/intra class interactions, we highlight in figure 5.8, the main classes driving the structure of the network. We present the results from the SBM on the classes with 50 authors or more. This reorganization clearly confirmed the presence of a clique of mainly international contributors who tend to collaborate rarily outside their class. The larger size here (Figure 5.8) is very diversed and contains all regional contributors to the malaria research effort. The presence of 3 smaller cliques which collaborate intensively between themselves is worth noting as well (See inter/intra class probabilities network on figure 5.8).

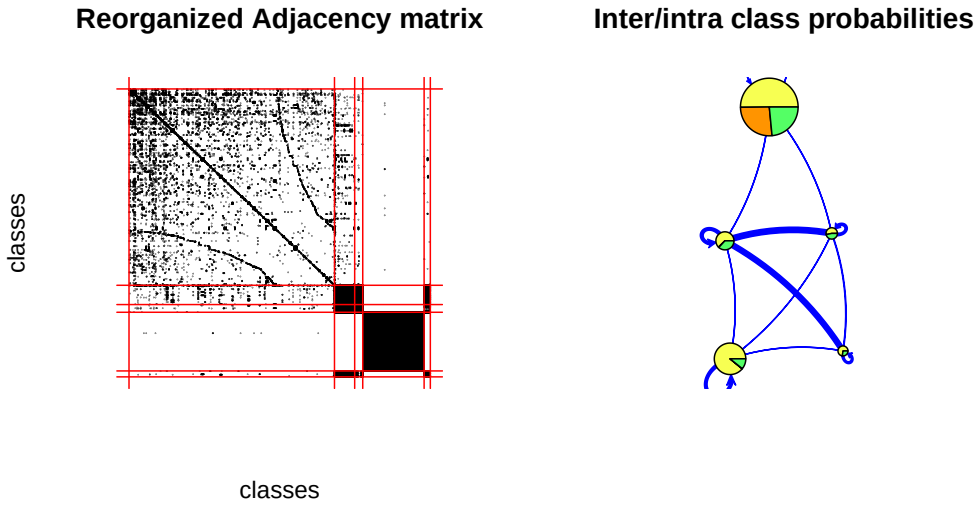


FIGURE 5.8: Summary of the goodness-of-fit of the SBM analysis highlighting interactions between the top 5 larger classes of the Malaria co-authorship network.

5.3.2.2 Exponential Random Graph Model

Table 5.3 summarizes the results of the different models we fit to the observed network. Model 1 is analogous to the null model in a typical General Linear Model (GLM). The probability of any two authors establishing a collaboration tie is therefore expressed as the inverse logit of the edge coefficient. The inverse logit of a coefficient x is defined as $logit^{-1}(x) = 1/(1 + exp(-x))$. The conditional log-odds for a collaboration between authors in the network is -2.76. The associated probability of any two authors establishing a collaboration tie is therefore 5.96%. To put this in perspective, this probabibility is the same as the density of the malaria co-authorship network. Since, our network is characterized by a high transitivity, we modeled the triangle ERGM term along with the edge term in model 2. We see some improvements in the model performance with a significantly positive but small triangle effect on the collaboration tie formation (Coefficient

$= 0.08$, $p < 0.001$).

In model 3, we describe the co-authorship network as a function of the number of collaborations, the number of publications, and the number of citations of authors inside the network. We also include confounding homophily term on cluster assignment from the SBM and on the collaboration type. Compared to models 1 and 2, model 3 has tremendously improved (See AIC and BIC in table 5.3). The edge effect has decreased (Coefficient $= -7.98$, $p < 0.001$) with the associated conditional probability (given all other terms in the model) equal to 0.03%. We observed a small, though positively significant effect of the number of collaborators and the number of publications on the odds of collaboration tie formation between any two authors. One unit increase in the number of collaborators increases the odds of collaboration tie by 2% while one unit increase in the number of publications increases the odds of establishing a collaboration tie by 12.75%. On the other hand, model 3 has found a very small but significant negative effect of the number of times an author was cited on the odds of collaboration tie formation. One unit increase in the number of citation of a given author was associated with 1% decrease in the odds of collaboration between two authors conditional on all the other terms in the model.

It clearly appears that the process underlying the malaria co-authorship network is driven by homophily on cluster assignment or membership to a specific research community and the type of collaboration. The conditional probability of two authors collaborating adjusted by the homophily on their membership to a research community is estimated at 8.32% compared to the baseline probability of 0.03% given all other terms in model 3.

Adjusted by the collaboration type, the same probability is estimated at 0.05% conditional on all other terms in the model. The overall conditional probability adjusting for all terms in model 3 is estimated at 14.06% which is a lot greater than the 5.95% estimated from model 1.

In model 4, we introduced factor attributes on the collaboration type in order to investigate the likelihood of researchers affiliated to Beninese institutions to establish international and regional or African collaboration ties. While model 4 slightly improved upon model 3, it displays minor changes in the coefficient of the terms it has in common with model 3. Overall, compared to researchers with international research affiliations, researchers affiliated to Beninese research institutions have 37.7% average decrease in the odds of establishing collaboration ties. On the other hand, researchers affiliated to other African research institutions have 78.6% increase in the odds of establishing a collaboration tie than researchers affiliated to international research institutions. In other words, in model 4, the probability for researchers affiliated to international institutions to establish a collaboration tie is estimated at 14.19%, that of researchers affiliated to Beninese institutions is 10.72%, and that of researchers affiliated to African institutions other than Beninese institutions is 22.79%.

None of the structural models containing high order ERGM terms, nor the models containing the dyadic attribute terms converged after the maximum of 1,000 iterations making estimates from these models unreliable. This observation justifies the reason why we do not present the results from these models in table 5.3. The unability of model containing structural terms to converge also makes it impossible for us to assess model degeneracy

as recommended by Handcock et al. [125].

TABLE 5.3. ERGM of the co-authorship Malaria network.

	Model 1	Model 2	Model 3	Model 4
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Network structural predictor				
Intercept(edge)	−2.76 (0.00)***	−5.00 (0.01)***	−7.98 (0.02)***	−8.22 (0.02)***
Triangle	–	0.08 (0.00)***	–	–
Number of collaborations	–	–	0.02 (0.00)***	0.01 (0.00)***
Number of publications	–	–	0.12 (0.00)***	0.13 (0.00)***
Number of times cited	–	–	−0.01 (0.00)***	−0.01 (0.00)***
Homophily on cluster assignment	–	–	5.58 (0.02)***	5.68 (0.02)***
Homophily on collaboration type	–	–	0.46 (0.01)***	0.61 (0.00)***
Factor attribute effect (collaboration type)				
International	–	–	–	*REF*
National	–	–	–	−0.32 (0.02)***
Regional	–	–	–	0.58 (0.01)***
Number of iterations	6	18	8	9
Akaike's Information Criterion (AIC)	725268	660444	220964	217026
Bayesian Information Criterion (BIC)	725280	660469	221038	217125
Model Log Likelihood	−362633 ($df = 1$)	−330220 ($df = 2$)	−110475.9 ($df = 6$)	−108505.2 ($df = 8$)

REF = reference, *SE* = Standard Error, *df* = degree of freedom
$^{***}p < .001$
$^{**}p < .01$
$^{*}p < .05$

Figure 5.9 presents the goodness-of fit of model 4. The observed properties are depicted by the black lines. Gray lines with circles represent the 95% confidence intervals for the simulated network properties. Goodness-of-fit is asserted when the black lines lie in-between the confidence intervals lines. The wide range of degree distribution of our co-authorship network makes it difficult to assess model fit in terms of degree distribution. But it is clear that in general, model 4 fits poorly to the observed network despite the highly significant estimates obtained. We therefore have strong evidence confirming that there is likely something other than the terms included in this model that are driving the structure of the network, possibly additional attributes our study did not control for. The following section attempts to address this shortcoming.

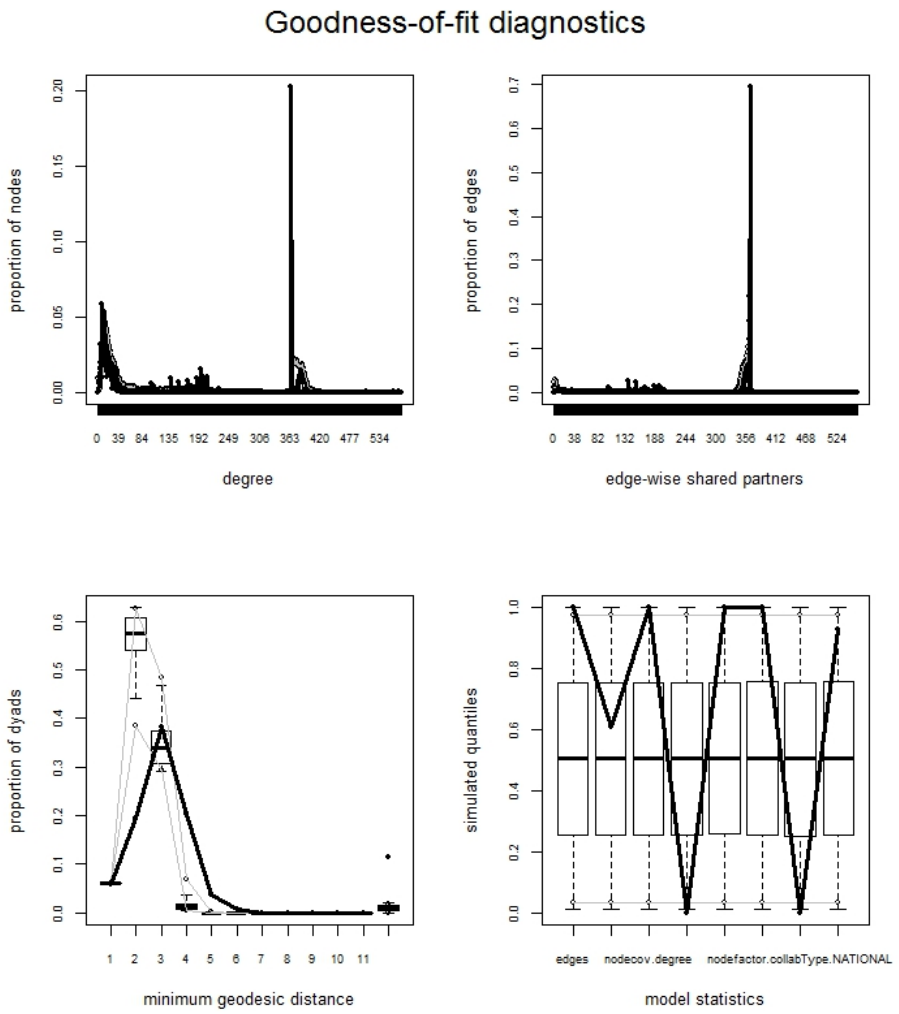


FIGURE 5.9: ERGM goodness-of-fit of final model 4 assessment.

5.3.2.3 Temporal Exponential Random Graph Model

The observed cumulative network was subset in seven snapshots representing respectively the following time spans: 1996 – 2006, 2007 – 2009, 2010 – 2011, 2012 – 2013, 2014, 2015 and 2016. Figure 5.10 displays the topological structure of the snapshots of the different time steps.

Table 5.4 summarizes the results of the different temporal models we fit to the observed

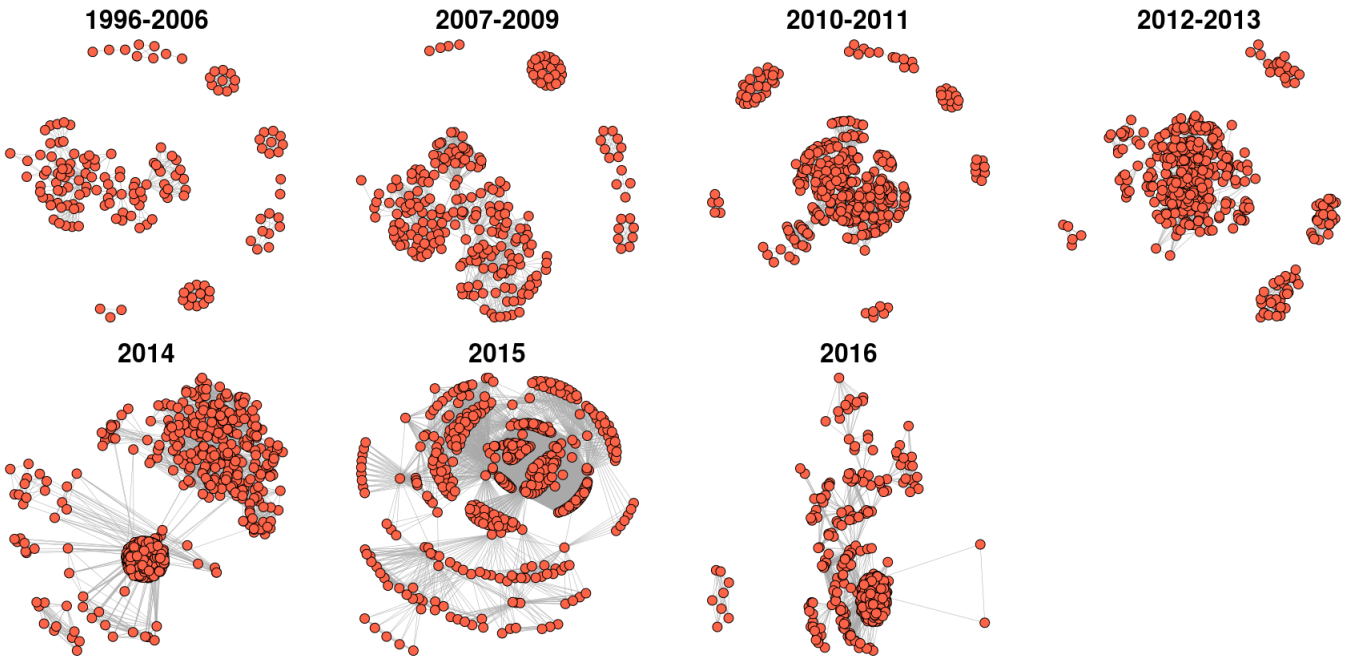


FIGURE 5.10: Topological structure of the different snapshots of the malaria co-authorship network.

network. Models 1, 2, and 3 are equivalent to a pooled ERGM across the 7 different time points (Fig. 5.10). The null model of the TERGM (model 1) suggests that the baseline log-odds for collaboration tie formation between authors in the network is -4.66. This coefficient is equivalent to a baseline probability of 0.9% for any two authors in the network to establish a stable collaboration tie. This probability is significantly lower than the 5.96% baseline probability of collaboration tie establishment reported by the ERGM (section 5.3.2.2).

Model 2 of the TERGM describes the co-authorship network as a function of the number of collaborations, the number of publications, and the number of citations of authors inside the network. It is also adjusted by homophily on cluster assignment from the SBM and on the collaboration type. Compared to model 1, model 2 has slightly improved (See AIC

and BIC in table 5.4). The edge effect has decreased (Coefficient $= -10.14$, $p < 0.001$) with the associated conditional probability (given all other terms in the model) equal to 0.004%. We observed a relatively high positively significant effect of the homophily on cluster assignment on the odds of collaboration tie formation between any two authors. Adjusting for the other variables in model 2, authors of the same research groups/communities are 4.96 times as likely to collaborate than authors that belong to different research groups. The effect of the other attributes in model 2 are minor. When we adjust for attribute effect on the collaboration type, we obtained model 3 which is slightly better than model 2. Relatively to model 2, the edge effect decreases more followed by an even stronger effect of the homophily on cluster assignment of the authors in the network (Coefficient $= 5.06$, $p < 0.001$).

After introducing temporal dependencies terms, we obtained model 4 which tremendously improved compared to models 1, 2 and 3. Model 4 confirms the observation made in section 5.3.2.2 that the process underlying the malaria co-authorship network is driven by homophily on cluster assignment or membership to a specific research community and the type of collaboration. It further confirms that the linear trend suspected observed in figure 5.1 is significantly associated with the odds of collaboration tie formation in the Malaria co-authorship network. Model 4 suggests that the baseline conditional probability of any two authors to collaborate is estimated at 0.02% given all other terms in the model. The coefficient associated to the dyadic stability term is 1.07 meaning that the odds of existent and non existent collaboration ties at one time point to remain the same at the next time point increased on average by 65.7%. In other words, the odds of

new collaboration ties and non-ties to occur from one time point to another is 34.3%. In addition, the TERGM showed that the probability of sustainable collaboration tie formation among international researchers is 12.13% versus 12.24% for researchers affiliated with national institutions ($p > 0.05$). However, this probability significantly increases to 20.26% for researchers affiliated to African research institutions other than those in Benin. These probabilities confirm the results from the ERGM final model with respect to the higher probability of tie formation between researchers affiliated to African institutions other than Beninese institutions. None of the structural temporal models containing high order TERGM terms, nor the models containing the dyadic attribute terms converged after the maximum of 1,000 iterations making estimates from these models untrustful.

TABLE 5.4. Temporal ERGM of Malaria co-authorship network.

	Model 1	Model 2	Model 3	Model 4
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Network structural predictor				
Intercept(edge)	−4.66 (0.00)***	−10.14 (0.02)***	−10.45 (0.02)***	−8.65 (0.05)***
Number of collaborations	–	0.03 (0.00)***	0.03 (0.00)***	0.03 (0.00)***
Number of times cited	–	−0.03 (0.00)***	−0.02 (0.00)***	−0.03 (0.00)***
Number of publications	–	0.45 (0.00)***	0.46 (0.00)***	0.45 (0.00)***
Homophily on cluster assignment	–	4.96 (0.02)***	5.06 (0.02)***	4.79 (0.02)***
Homophily on collaboration type	–	0.44 (0.01)***	0.56 (0.01)***	0.54 (0.01)***
Factor attribute effect (collaboration type)				
International	–	–	*REF*	*REF*
National	–	–	−0.10 (0.02)***	0.01 (0.02)
Regional	–	–	0.55 (0.01)***	0.60 (0.01)***
Temporal dependencies				
Dyadic stability	–	–	–	1.07 (0.01)***
Linear trends	–	–	–	−0.18 (0.01)***
Akaike's Information Criterion (AIC)	94681198	93740511	93737596	67005816
Bayesian Information Criterion (BIC)	94681230	93740624	93737742	67005991
Model Log Likelihood	−47340597	−46870248	−46868789	−33502897

REF = reference, *SE* = Standard Error
***$p < .001$
**$p < .01$
*$p < .05$

Figure 5.11 presents the goodness-of-fit assessment for the TERGM model 4. We can see that this model containing temporal dependencies fits better to the observed Malaria co-authorship network than the final ERGM model 4. While the first five subfigures compare the distribution of endogenous network statistics between the observed network and the simulated ones, the last subfigure presents the Receiver Operating Characteristics (ROC) and precision-recall (PR) curves. In general, the closer the curve is to the left-hand border and the top border of the ROC space, the more accurate the prediction is. On the other hand, the closer the curve is to the 45-degree diagonal of the ROC space, the less accurate is the prediction. The ROC for model 4 is depicted by the dark red curve compared to the ROC of a random graph depicted by the light red curve. Similarly, the dark blue curve represents the PR of model 4 versus the light blue curve representing the PR of a random graph [113]. It clearly appears that the final TERGM model 4 outperformed the random null model with an Area Under the Curve (AUC) value estimated at 79.98%.

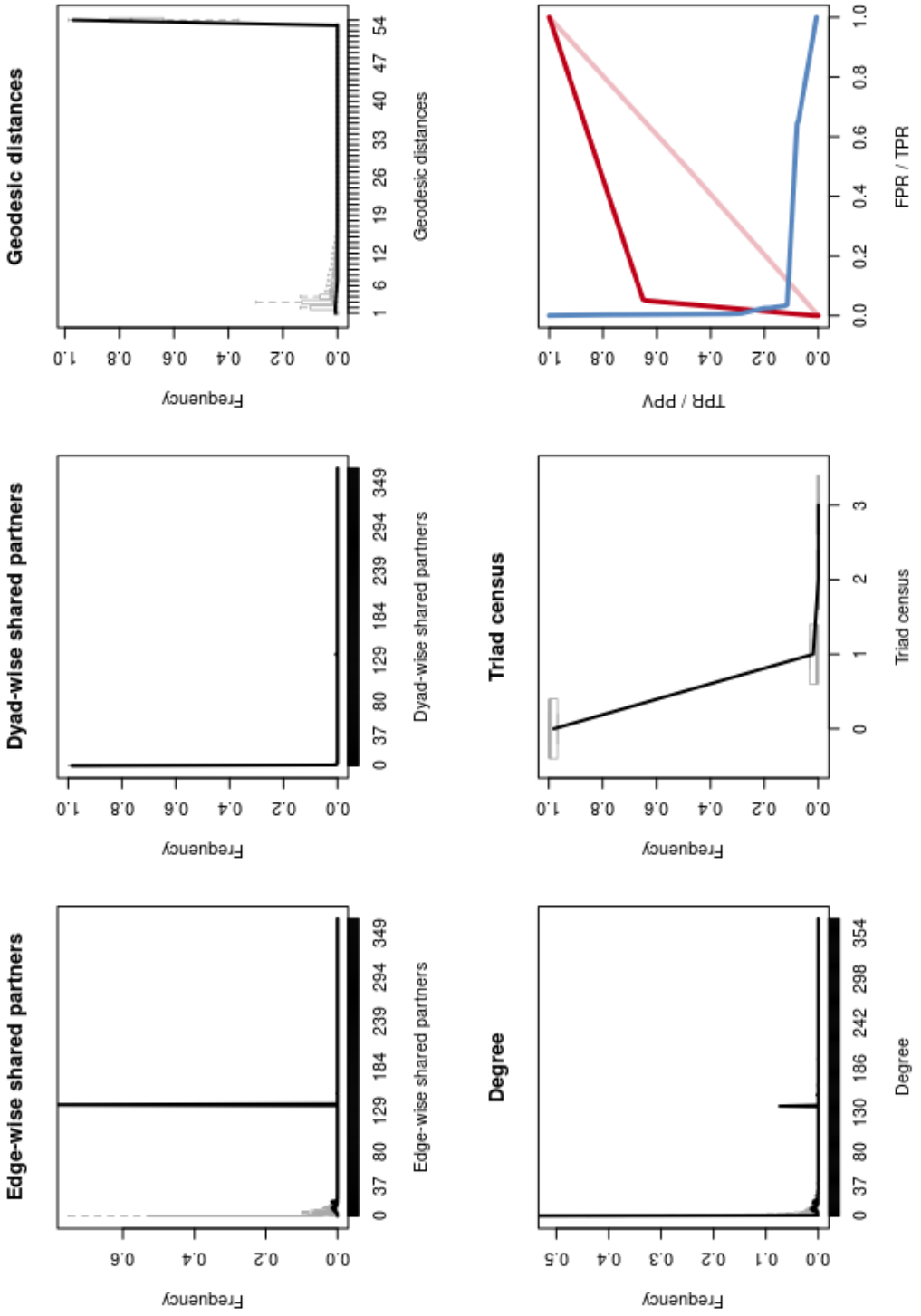


FIGURE 5.11: Goodness-of-fit assessment for the final Malaria TERGM Model 4 with temporal dependencies.

5.3.2.4 Latent Network Model

Figure 5.12 presents a 3-dimensional visualization of the Malaria co-authorship network, with layouts determined according to the inferred latent eigenvectors from the no pair-specific model (on top), the model containing nodal covariates (middle), and the model containing nodal and dyadic covariates (bottom). Blue vertices represent authors affiliated to Beninese research institutions, Red vertices are authors affiliated to international institutions, Gold vertices represent authors affiliated to African research institutions other than Benin, and White vertices represent authors with no determined affiliations. Node sizes are proportional to the betweenness value of each vertex. Looking at the three visualizations, it clearly appears that the first two visualizations are somewhat similar while the third is different. In fact, in the first two visualizations, the authors are clustered in mainly three clusters. We can see that all the authors affiliated to Beninese research institutions (in blue) are clustered in one cluster while authors with international affiliations (in red) and regional authors (in gold) are distributed across all three main clusters. These observations suggest a significant geography effect on the odds of collaboration tie establishment in the malaria co-authorship network.

The first two visualizations also highlight key brokers that liaison between clusters. In the third visualization, on the other hand, there appears to be only one main cluster. This last observation suggests that the nodal covariates and mainly homophily on research community membership and type of affiliation explain much less coarse-scale network compared to dyadic covariates. Indeed, when the dyadic covariates are added to the model, there

is less structure left to be captured by the latent variables. These results compensate the lack of-fit of the ERGM model and confirmed our findings in the previous section.

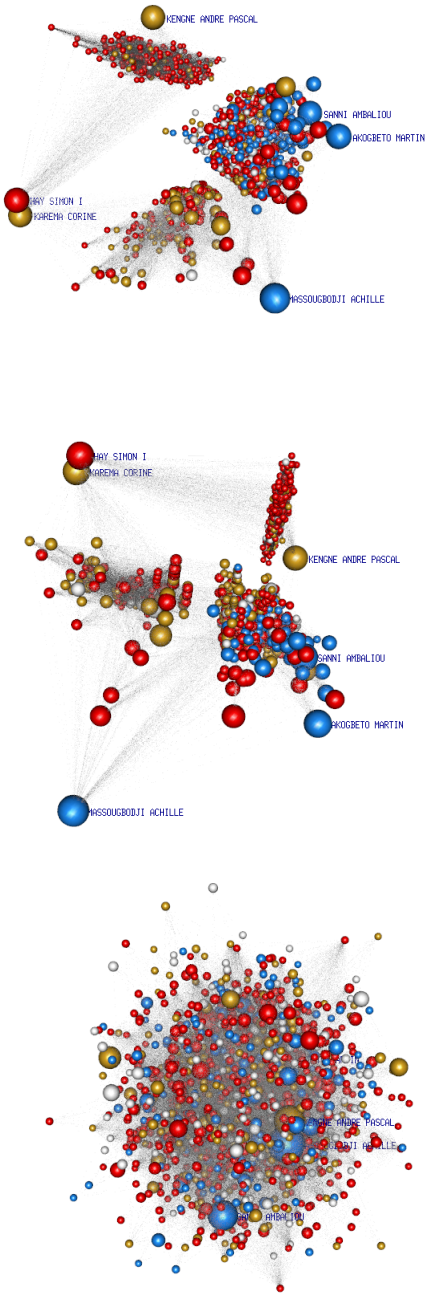


FIGURE 5.12: Visualizations of the Malaria co-authorship network with layouts determined according to the inferred latent eigenvectors in the LNM models (International (Red); Regional (Gold); Local (Blue)).

The ROC curves on figure 5.13 show that the first two models appear to be comparable in their performance from the perspective of edge status prediction with an Area Under the Curve (AUC) being roughly 98.8%.

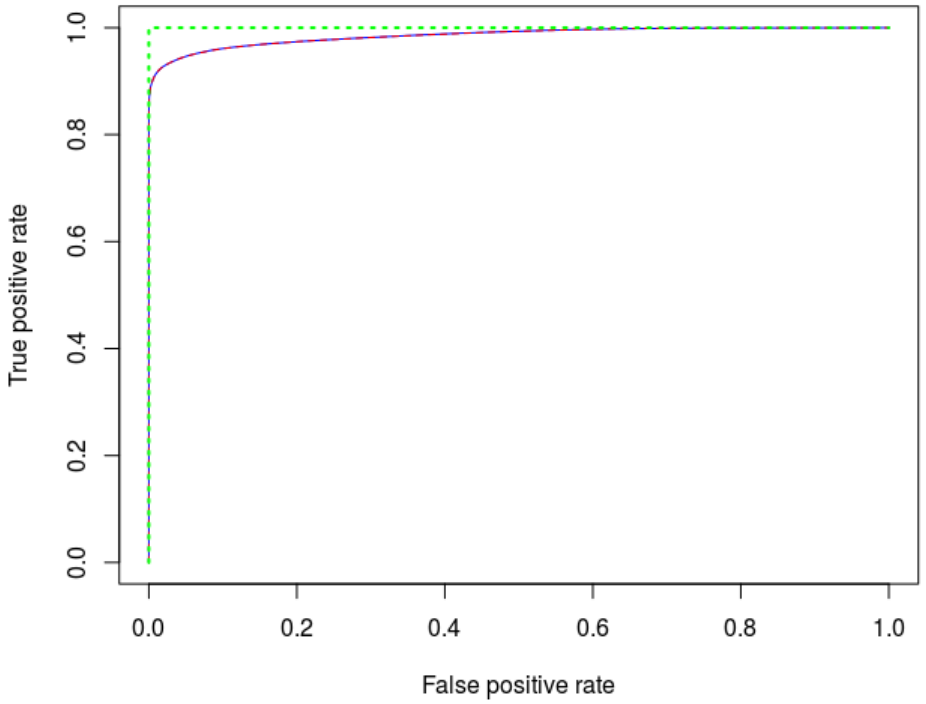


FIGURE 5.13: ROC curves comparing the goodness-of fit of the Malaria co-authorship network for three different eigenmodels, specifying (i) no pair specific covariates (blue), (ii) nodal covariates (red), and (iii) nodal and dyadic covariates (green), respectively.

5.4 Discussion and Conclusion

In this chapter, we provide insights in the structural characteristics of the malaria co-authorship network in the Republic of Benin over a relatively long period. The 20 years of data collected coincides with the onset of active malaria research from 1996 until today. The significant increase in malaria research and collaborations (figure 5.7) between the authors over the years is an expected finding given the regain and renewed interest in malaria control and elimination goals set forth [126, 127]. Our results show that the mechanism underlying the formation of the malaria co-authorship network in Benin is not random. It further demonstrates that the malaria research collaboration network in Benin is a complex network that seems to display small-world properties (often referred to as "six degrees of separation").

The non-trivial number of authors with higher order of magnitudes confirms the presence of closed research groups where collaborative research likely happens only among members. In other words, interdisciplinary collaboration tends to occur at higher levels between prolific researchers with the majority of the collaborations happening between researchers from the same scientific communities. Prominent authors with long tenure tend to collaborate with similar authors, young or less prolific authors tend to collaborate with both prolific authors and authors with very few collaborations. Similar findings were reported by Janet Okamoto [128] who studied scientific collaboration on a much smaller scale. Key brokers facilitate scientific collaborations within and outside their scientific

community [68]. Betweenness centrality measures identify such brokers who are important hubs for inter and transdisciplinary research. Many of the main brokers proved to also be the most connected and the most central authors confirming the presence of long publishing tenure authors in our network [129]. The flow of information in this network in Benin is slow as it only relies on 16 authors representing less than 1% of all the authors in the network. Such a low information flow was also reported by Salamatia and Soheili [70] in a 2016 study on a co-authorship analysis of Iranian researchers in the field of violence. Generally, the most important authors in a co-authorship network are the ones with the highest degree of collaborations [130, 131]. However, to the long-term substainability of the malaria research network in Benin, the 16 authors identified as cut vertices are the most important authors. In other words, the removal of less than 1% of the authors from the network would lead to its collapse. Such a collapse would undoubtedly be detrimental to the future of malaria research in Benin. This finding clearly confirms the conclusion of Toivanen and Ponomariov [67] that the African research collaboration network is vulnerable to structural weaknesses and uneven integration.

Small-world networks are known to have small shortest path distance and a high clustering coefficient. Although this co-authorship network seems to display such properties, the Monte-Carlo simulations revealed that the observed network has unexpected properties compared to classic small-world networks. A study of co-authorship network conducted on Chagas disease has found similar findings [26]. Unlike our study, the authors of the Chagas disease co-authorship study did not deepen their analysis to confirm the small-world nature of their observed network. Other mechanisms such as preferential attachment

have been found to explain the structure of international scientific collaboration network [132]. Unlike those studies, our network displayed unexpected properties that are more extreme than the 4 mathematical models we simulated. Our network has significantly higher clustering than expected from the 4 mathematical models presented here. One observation we are sure of is that none of the random graph models used here tend to explain the growth and the structure of the malaria co-authorship network in Benin. We therefore claim without any doubt that the structure and growth of our network is not random confirming the presence of hidden factors explaining the current structure of the network. Assessing such factors and the extent to which they influence scientific collaborations is important for the future of malaria research and its long-term sustainability. Unfortunately, none of the proposed mathematical models seem to accurately describe the observed structure of the network. To address these limitations, advanced statistical modeling was used to further explain the structure of the network.

Our first approach to modeling our network relied on the use of SBM. In addition of being a model based clustering method, the SBM identified important organizational and interactional patterns in the network. It identified a large clique of mainly international researchers with little or no collaborations with other research groups. It also identified the main broker authors in the network. For example, in the first two visualizations on figure 5.12, the brokers with affiliations to national institutions are MASSOUGBODJI ACHILLE, AKOGBETO MARTIN, and SANNI AMBALIOU. These authors are also the ones with the highest citation counts. Such an observation is not surprising given their long tenure, their publication records, and known expertise in malariology, parasitology

and medical entomology. The overwelming dominance of regional and international players in the network is consistent with previous observations by Onyancha and Maluleka [133] who concluded on a much higher likelihood of Sub-Saharan African countries to collaborate with non-African states.

Overall, the ERGM and TERGM show that the mechanistic phenomenon driving collaboration ties in the malaria research in Benin is influenced by homophily on the type of affiliation (national, international or regional) and on membership to a research group or cluster, verifying therefore our third hypothesis. The models clearly show that the dominance of the Beninese malaria research arena by international and regional players, and further demonstrates the lower likelihood of local Beninese researchers to establish international collaboration ties compared to regional researchers. This latter finding has been confirmed by the LNM which also confirms our second hypothesis. The ERGM and the TERGM revealed that factors such as number of publications, number of citations and number of collaborations are associated to higher likelihood to establishing collaboration ties, confirming therefore our first hypothesis.

It is worth noting that many of the studies on co-authorship network analysis are descriptive in nature. This study is one of the rare co-authorship network analysis to model a co-authorship network using advanced statistical models. ERGM is the leading approach to modeling network [134]. The literature has reported application of this model in studying various social network such as the analysis of friendship and obesity [135, 136], the exploration of the association between hormone and social network structure [137]. Similarly to friendship networks, the use of ERGM to model co-authorship networks is easily

justified. However, the size of our network prevented the fitting of complex models including dyadic and structural terms. In addition, our best ERGM model failed to adequately fit the observed network data. This lack of goodness-of fit, according to Hunter, Goudreau and Handcock [138], could be improved by including the geometrically weighted edgewise shared partner, geometrically weighted dyadic shared partner, and geometrically weighted degree network statistics to our model. Although, we follow such recommendations by including these structural network statistics to our final model, the ERGM model failed to converge after a maximum of 1,000 iterations. At about 750 iterations, we noticed that the processing became both computationally intensive and expensive in terms of CPU time and memory usage. In a recently published paper, Schmid and Desmarais [134] acknowledged the difficulty of fitting network which size is of the order 1,000 vertices using ERGM. They recommended that using the maximum pseudolikelihood estimation (MPLE) instead of the Monte Carlo maximum likelihood (MCMLE) could tremendously reduce computation time. Having followed these recommendations too, the ERGM model containing dyadic and structural terms still failed to converge. By finally including temporal dependencies and fitting a temporal ERGM, we have tremendously improved the fitness with a predictive performance of roughly 80%. Nevertheless, we suspect that the number of edges, the large size of the network added to the possibility of hidden/latent variables might justify the failure of the models containing the dyadic and structural endogenous terms to converge. We remedy this situation by applying LNM to the observed network data.

All three latent network models (LNM) proved to be successful in fitting the observed

network data. A study by Kronegger et al. [139] conducted an investigation aiming at describing the collaboration in Slovenian scientific communities using data from four different disciplines. Their methodological approach is consistent with ours. The main difference is their application of Stochastic Actor-Oriented Model (SAOM) on the dynamics of their co-authorship networks. Since the SAOM is an actor-oriented modeling method and we are interesting in tie prediction here, we relied rather on a tie-oriented approach by applying the TERGM to our network data.

Our results suggest that the regain in Malaria research funding has appealed to research groups all around the world, hence the explosion in publications number and research collaborations. As the disease continues to be main public health concern in the Republic of Benin, it is essential to consolidate the knowledge generated from the numerous studies on the disease and reinforce the different communities involved in the research effort. In addition, there is an urgent need to reinforce the malaria research network in Benin by continuously supporting, stabilizing the identified key brokers and most productive authors, and promoting the junior scientists in the field. However, we observed a tendency of the international researchers to only collaborate among themselves. Although the rise in scientific collaboration between advanced and developing nations [140], the latter observation may limit effective and sustainable technology transfer in Benin. It is possible that some of the isolated cliques within the network have top-notch research capabilities and skills researchers affiliated to Beninese institutions can acquire, should the research groups be more inclusive. Unfortunately, our visualizations showed that broker authors

that liaison those closed groups to national researchers tend to be regional or international researchers as well. We therefore recommend, that policies should be designed, at international, regional and country level, to diversify research groups operating in any Sub-Saharan African countries. Such policies will ultimately enable effective technology transfer, multidisciplinarity, and promote junior African researchers to advance the search of a solution to the Malaria problem in Africa and particularly, in Benin.

Chapter 6

Results: The HIV/AIDS Co-authorship Network

6.1 Data

The literature search was conducted in the Web Of Science (WOS) using combinations of HIV/AIDS related MeSH terms including ”HIV”, ”AIDS”, ”VIH” and ”HIV Infections”. The final query set (Table 6.1) returned 237 records. After a rigorous screening process, 102 documents met the selection criteria. On average, there were 9.47 authors per published document.

The Author Name Disambiguation process led to the identification of 516 unique authors with a precision of 99.88% and a recall of 82.54%. The generated multigraph co-authorship network therefore contained 516 vertices (authors) and 5,114 parallel edges

TABLE 6.1. HIV/AIDS Bibliographic Search Queries.

Set	Queries	Results
#1	TOPIC: (HIV AIDS) Refined by: COUNTRIES/TERRITORIES: (BENIN)	52
#2	TOPIC: (HIV AIDS) AND ADDRESS: (BENIN)	107
#3	TOPIC: (HIV) OR TOPIC: (AIDS) AND ADDRESS: (BENIN), Refined by:COUNTRIES/TERRITORIES: (BENIN)	182
#4	TOPIC: (HIV) OR TOPIC: (VIH) OR TOPIC: (AIDS) AND ADDRESS: (BENIN), Refined by: COUNTRIES/TERRITORIES: (BENIN)	182
Final Set	#1 OR #2 OR #3 OR #4	237

(collaborations). The number of unique authors for HIV research is roughly one third of the Malaria ones. As displayed in figure 6.1, we can see the significant increase in publications, scientific collaborations and the number of authors involved in HIV/AIDS research from 2008 until 2016. This general upward trend seems to be linear from the year 2008 to 2016. The variation seen between adjacent years may reflect the relatively small productivity of HIV research prior to the year 2008.

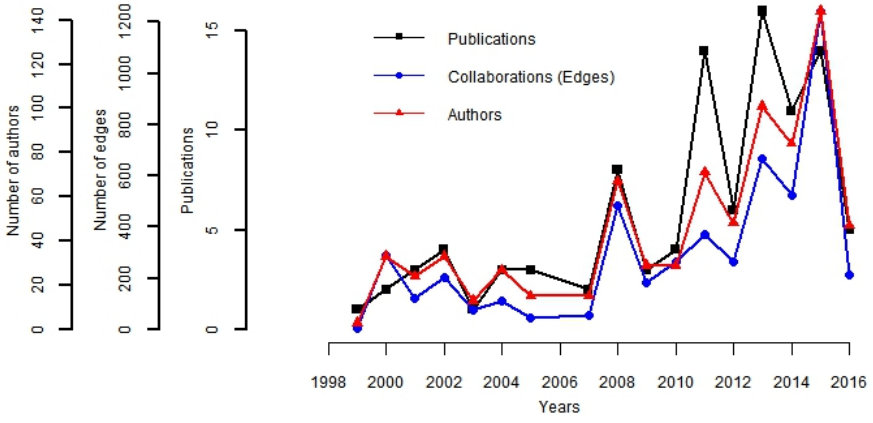


FIGURE 6.1: Evolution of the published HIV related documents, authors and collaborations from January 1996 to December 2016

6.2 Descriptive Data Analysis

For the multigraph network, the degree distribution varies between 1 and 403 with an average degree distribution of 19.82 and a median of 12. In addition, there was a substantial number of vertices with low degrees (Fig. 6.2). The log scale distribution of the degrees on figure 6.3 reveals that there was also a non-trivial number of vertices with higher order of degree magnitudes. There is a tendency of the vertex degrees to follow a heavy-tail distribution suspected on figure 6.2.

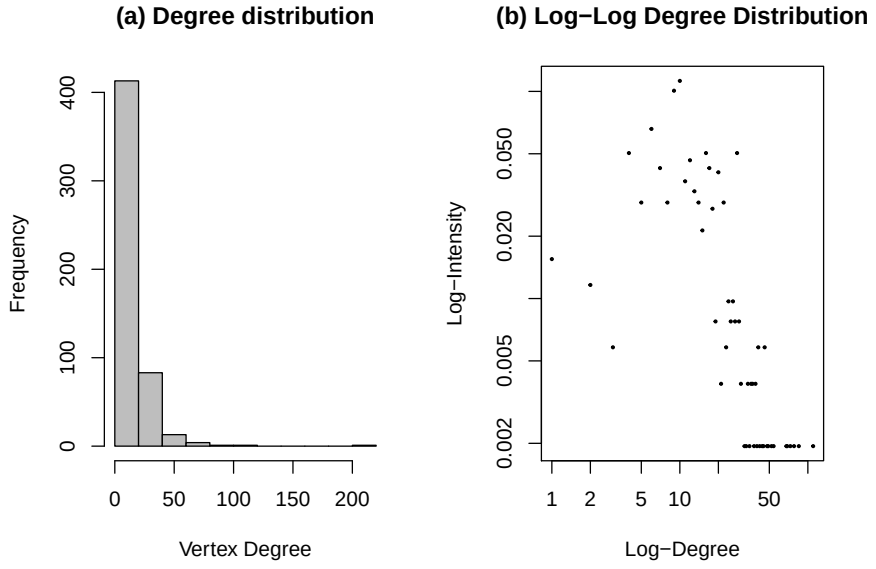


FIGURE 6.2: Degree distribution of the HIV/AIDS co-authorship network

After we convert the multigraph network in a weighted graph, it results in a simple graph of 516 vertices and 3,966 weighted edges. Closeness centrality measures range between 3.76×10^{-6} and 3.19×10^{-5} with a median of 3.13×10^{-5}. Betweenness measures range between 0 and 49,280 with a median of 426.2. A network visualization with the vertices' size proportional to betweenness centrality measures clearly reveals the presence of broker

authors (Figure 6.4 and Table 6.2). The median Eigenvectors is 0.202 with a mean of 0.045. The eigenvectors measures confirm the presence of author hubs in the network suggesting the presence of closed collaboration groups. Table 6.2 presents a list of the 10 author hubs with the highest Eigenvectors values.

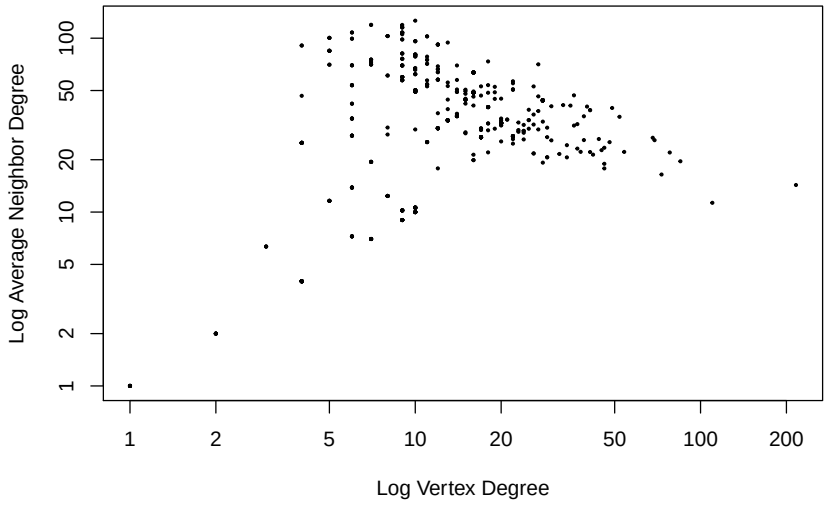


FIGURE 6.3: Log-Average Neighbor degree Distribution of the HIV/AIDS co-authorship network

Edge betweenness centrality measures identify co-authorship collaboration ties that are important for the flow of information. Table 6.2 presents the top 10 most important collaboration ties for the flow of information in the HIV/AIDS co-authorship network in Benin.

TABLE 6.2. List of the most important authors and collaborations in the HIV/AIDS co-authorship network

Top 10 Brokers
ZANNOU DJIMON MARCEL
ALARY MICHEL
LEROY VALERIANE
AZONDEKON ALAIN
ANAGOUNOU SEVERIN
ADE GABRIEL
AZONKOUANOU ANGELE
NDOYE IBRA
NDOUR MARGUERITE
AFFOLABI D
Top 10 most connected authors (Top 10 network hubs)
ZANNOU DJIMON MARCEL
ALARY MICHEL
ANAGOUNOU SEVERIN
LOWNDES CATHERINE M
LABBE ANNIECLAUDE
DABIS FRANCOIS
MINANI ISAAC
BEHANZIN LUC
DIABATE SOULEYMANE
EKOUEVI DIDIER K
Top 10 most important edges for information flow
ZANNOU DJIMON MARCEL – LEROY VALERIANE
ZANNOU DJIMON MARCEL – NDOUR MARGUERITE
ALARY MICHEL – AZONKOUANOU ANGELE
ZANNOU DJIMON MARCEL – NDOYE IBRA
ANAGOUNOU SEVERIN – ADE GABRIEL
ZANNOU DJIMON MARCEL – WACHINOU ABLO PRUDENCE
ZANNOU DJIMON MARCEL – DALMEIDA MARCELLINE
AZONDEKON ALAIN – ADE GABRIEL
AZONKOUANOU ANGELE – AZONDEKON ALAIN
ZANNOU DJIMON MARCEL – COFFIE PATRICK A
Weak articulation points
ATADOKPEDE FELIX
NDOUR MARGUERITE
DALMEIDA MARCELLINE
AZONDEKON ALAIN
GANDAHO PROSPER
AFFOLABI D
ADE GABRIEL
ZANNOU DJIMON MARCEL

6.2.1 Network Cohesion

In total, 29 maximal cliques were detected in the network among which 2 cliques of size 24, 1 clique of size 23 and 4 cliques of size 3. Larger maximal cliques sizes range from 14 authors to 25 authors.

The HIV/AIDS co-authorship network has a density of 0.0298 indicating that the baseline probability of collaboration tie formation is 2.98%. The network also has a transitivity of 0.482 meaning that 48.2% of the connected triples in the network are close to form triangles. The transitivity metrics is a measure of the global clustering of the network.

The network is not connected and a census of all the connected components within the network reveals the existence of a giant component that dominates all the other connected components. The giant component of the HIV/AIDS co-authorship network includes 88.6% (457 vertices) of all the vertices in the network with the other components alone carrying less than 1% of the vertices (Fig. 6.4).

Information flow assessment of this network via cut vertices confirms the existence of 8 authors as the most vulnerable vertices in the network. Table 6.2 lists the authors that constitute the weak articulation points in the HIV/AIDS co-authorship network. The identification of cut vertices is a measure of the vulnerability of the HIV/AIDS co-authorship network [98].

Via the agglomerative hierarchical clustering method, we identify 24 different research communities (or clusters) which sizes range between 1 and 108 authors. Large research communities contain between 71 and 108 authors. Medium size research communities contain between 10 and 55 authors. Out of the 24 clusters detected, 12 are part of the

giant component. Figure 6.4 displays the structure of the network with each different colors representing each of the 24 clusters.

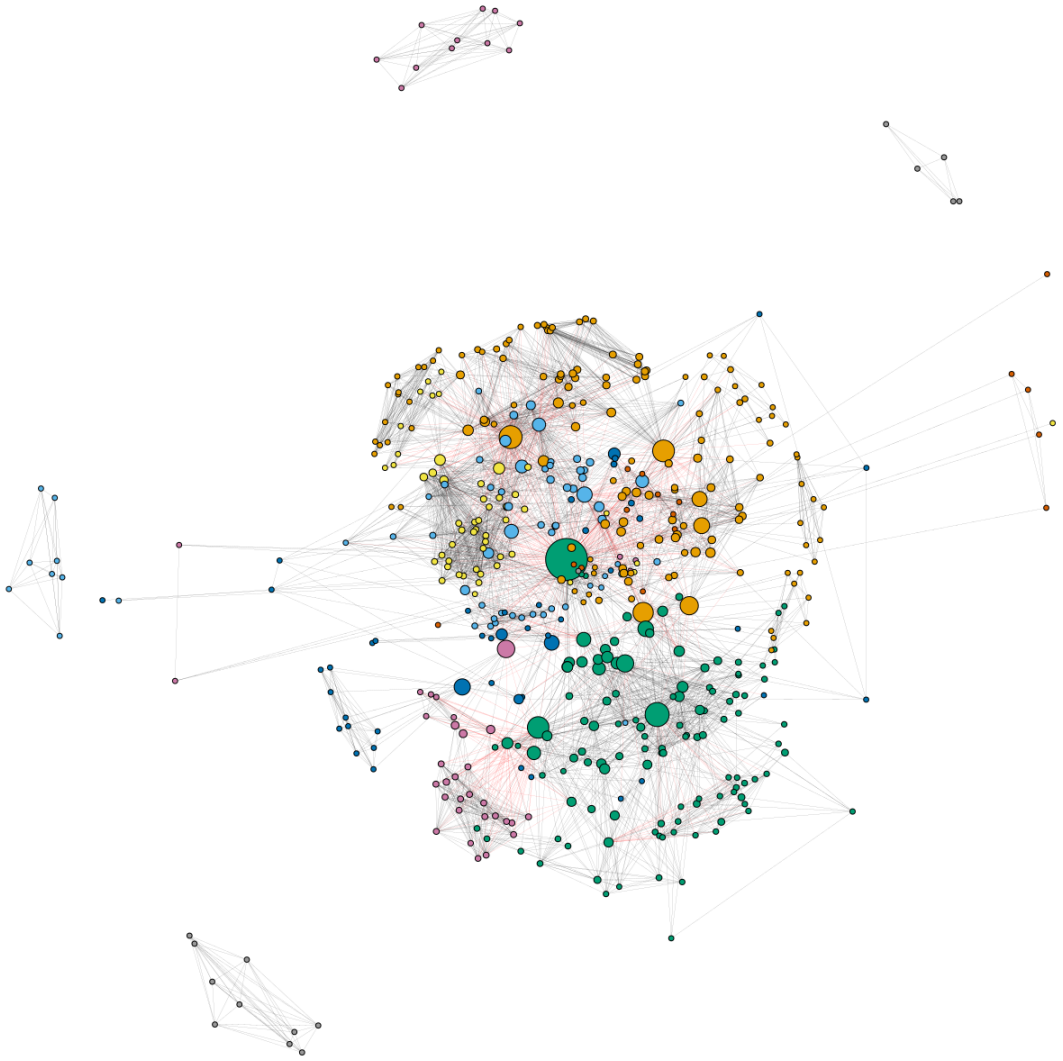


FIGURE 6.4: Topological Structure of the HIV/AIDS co-authorship network. Authors (vertices) of the same color belong to the same research community or cluster

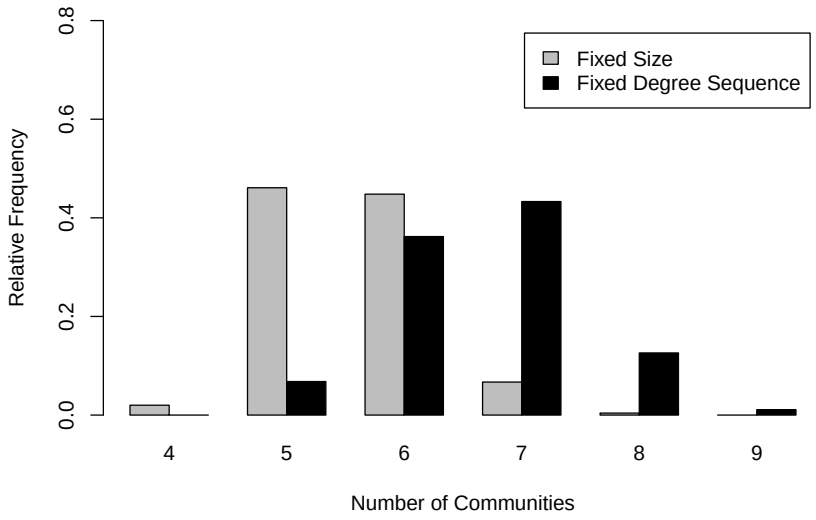


FIGURE 6.5: Monte-Carlo simulations of the HIV/AIDS network: Number of detected communities by the random graph models

6.3 Modeling

6.3.1 Mathematical Modeling

We performed 1,000 Monte Carlo based simulations to test the significance of the observed characteristics of the HIV/AIDS co-authorship network. Figure 6.5 clearly demonstrates that the number of communities detected is unusual from the perspective of both Classical random graphs and generalized random graphs (p-value < 0.0001). From the Classical random graph model, the expected number of communities was 5.574 (95%CI: 5.53 – 5.62). Similarly, the expected number of communities from the generalized random graph model is 6.65 (95%CI: 6.60 – 6.70).

Figure 6.6 displays the number of detected research communities using the Barabási-Albert's preferential attachment and the Watts-Strogatz models. The observed number of communities was extreme per both models (p-value < 0.0001). The expected number from the Watts-Strogatz model simulations is 3.181 (95%CI: 3.16 – 3.21) and 22.8 (95%CI: 22.7 – 23.0) from the Barabási-Albert model simulations. We also compared the clustering coefficient and the average shortest-path length. Let's recall that the observed clustering coefficient is 0.482. On one hand, there was substantially more clustering in our HIV/AIDS co-authorship network than expected from both random graph models (p-value < 0.0001). The expected clustering coefficients was 0.0365 (95%CI: 0.0363 – 0.0365) and 0.0842 (95%CI: 0.0841 – 0.0843) respectively for the classic random graph and the generalized random graph models.

On the other hand, there was substantially less clustering in our HIV/AIDS co-authorship network than expected by the Watts-Strogatz Small World model which expected clustering was 0.72615 (95%CI: 0.72611 – 0.72618).

We observed an average shortest-path length of 2.75 in the HIV/AIDS co-authorship network. This observed shortest-path length is significantly larger than what was expected from the random graph models (p-value < 0.0001) and significantly lower than what was expected from Watts-Strogatz small world model and the Barabási-Albert preferential attachment model (p-value < 0.0001).

The average shortest-path length was 2.49069 (95%CI: 2.49062 – 2.49077) and 2.381 (95%CI: 2.380 2.381) respectively for the classic random graph and the generalized random graph models.

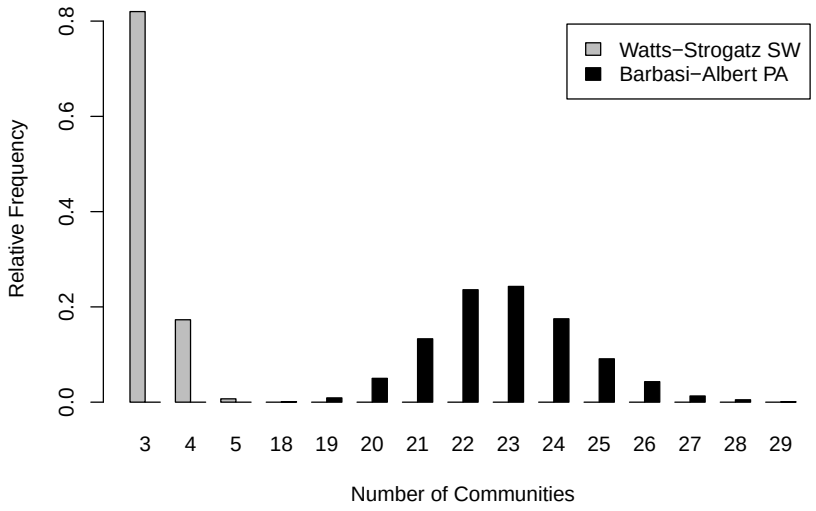


FIGURE 6.6: Monte-Carlo simulations of the HIV/AIDS network: Number of detected communities by the Watts-Strogatz and the Barabási-Albert models

For the Watts-Strogatz small world and the Barabási-Albert preferential attachment models, the average shortest-path length is respectively 5.31 (95%CI: 5.28 – 5.36) and 7.35 (95%CI: 7.31 – 7.38).

We performed the same simulations on the giant component of the network with similar results leading to similar outcomes.

6.3.2 Statistical Modeling

6.3.2.1 Stochastic Block Model

The SBM identifies 26 classes with a degree of latitude of 17 to 26 classes being reasonable (See ICL plot on figure 6.7).

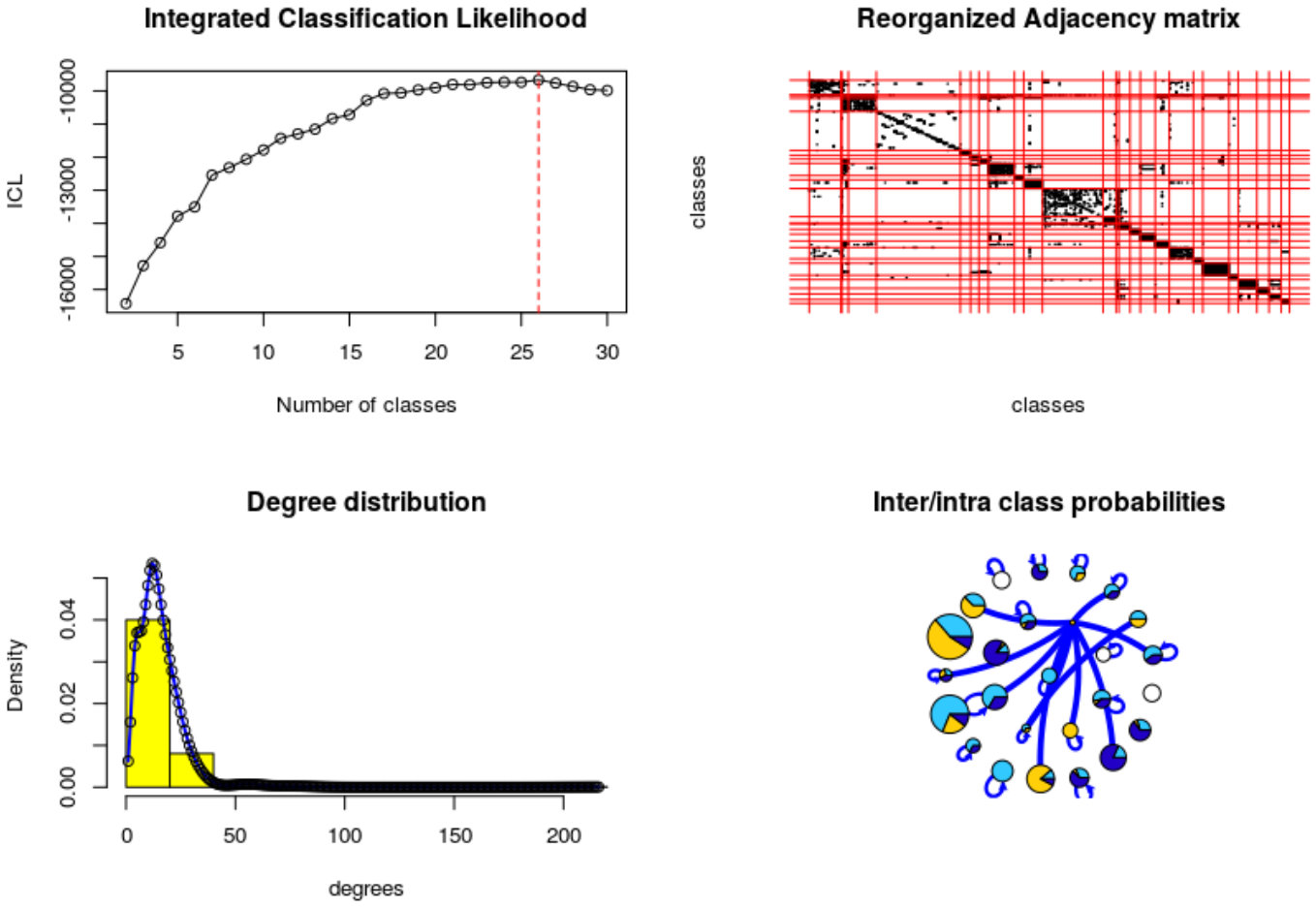


FIGURE 6.7: Summary of the goodness-of-fit of the SBM analysis on the HIV/AIDS co-authorship network.

Regarding the degree distribution, the fitted SBM describes well the observed degree distribution. On the network depicting the inter/intra probabilities between the classes, the

vertices represent the 26 identified classes, with each one of them divided into a pie chart displaying the proportion of authors of international affiliations (lightblue), authors of regional or other African affiliations (darkblue), and authors affiliated to Beninese research institutions (yellow). Generally, the dominance across the classes of international and regional players is observed. In addition, we observe denser ties between medium size and smaller size classes.

A close examination of the pie charts reveals that almost all the classes are heterogeneous. We note the presence of 2 large classes which are classes 5 and 12 (See reorganized adjacency matrix on figure 6.7). Class 5 is dominated by researchers with Beninese affiliations but appears sparser than class 12 which is dominated by international authors (Figure 6.7).

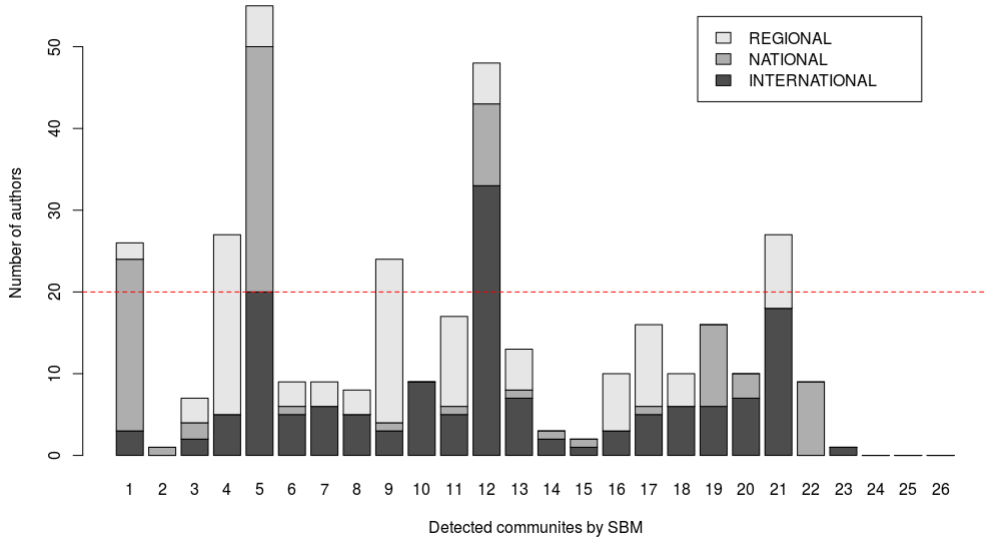


FIGURE 6.8: Distribution of national, international and regional authors by communities detected by the SBM in the HIV/AIDS network.

On figure 6.9, we present the SBM results emphasizing the largest classes (with more than 20 members). Here, we can confirm that smaller classes tend to collaborate more among themselves and intra-class collaborations tend to occur more.

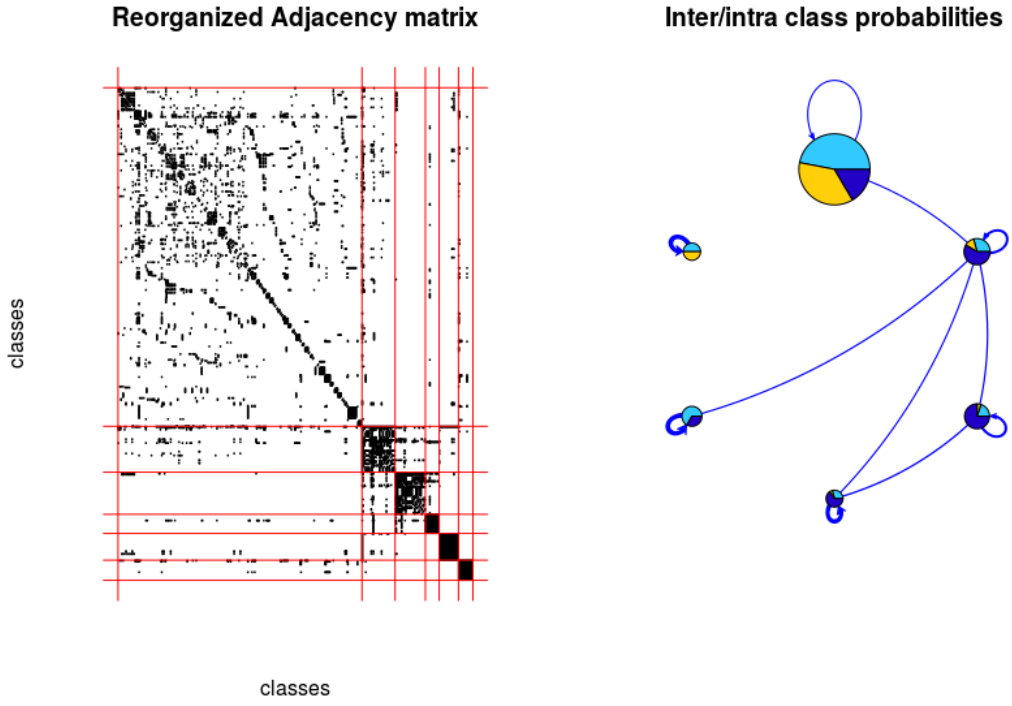


FIGURE 6.9: Summary of the goodness-of-fit of the SBM analysis highlighting interactions between the largest classes of the HIV/AIDS co-authorship network.

	Model 1	Model 2	Model 3
	Estimate (SE)	Estimate (SE)	Estimate (SE)
Network structural predictor			
Intercept(edge)	-3.48 (0.02)***	-7.51 (0.06)***	-7.55 (0.07)***
Number of times cited	–	0.00 (0.00)***	0.00 (0.00)***
Number of collaborations	–	0.08 (0.00)***	0.08 (0.00)***
Number of publications	–	-0.29 (0.01)***	-0.28 (0.01)***
Homophily on cluster assignment	–	5.01 (0.05)***	5.02 (0.05)***
Homophily on collaboration type	–	0.77 (0.05)***	0.72 (0.05)***
Factor attribute effect (collaboration type)			
International	–	–	*REF*
National	–	–	-0.05 (0.04)
Regional	–	–	0.21 (0.03)***
AIC	35668.54	18956.20	18912.74
BIC	35678.34	19014.98	18991.12
Log Likelihood	-17833.27	-9472.10	-9448.37

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 6.3. ERGM of the HIV/AIDS co-authorship network.

6.3.2.2 Exponential Random Graph Model

Different models were fit with the ERGM method (Table 6.3). Model 1, the null model, contains only the "edge" term. The inverse logit of the coefficient associated with this term is 0.0298 which is the baseline probability of collaboration ties establishment and also the density of the HIV/AIDS co-authorship network.

In model 2, we included all nodal variables, a homophily term on collaboration type and on cluster assignment determined from the SBM. Model 2 improved tremendously compared to model 1 (See AIC, BIC and model likelihood in table 6.3). We note a decrease

in the edge effect (Coefficient = -7.51, $p < 0.001$) with the associated conditional probability (given all the other terms in the model) estimated at 0.05%. For the remaining terms in model 2, we observed a positive and significant effect except for the number of publications. Model 3 differs from model 2 in that it includes a factor term on the collaboration type with a substantial improvement compared to model 2. Model 3 is therefore chosen as our last model. Regarding the number of publication, a one unit increase in the number of publication is associated with 32.3% average decrease in the odds of collaboration ties establishment. Model 3 further proves that the process underlying the structure of the HIV/AIDS co-authorship network in Benin is mainly driven by homophily on cluster assignment or membership to a research community or group (Coefficient = 5.02, $p < 0.001$). The conditional probability of any two authors belonging to the same research group to collaborate is estimated at 7.38% compared to the baseline probability of 2.98%. The same probability changes to 14.06% after adjustment by the collaboration type, and 11.82% after adjusting for the number of citations, collaborations and publications. Compared to researchers affiliated to international institutions, researchers affiliated to Beninese institutions have 5.1% average decrease in the odds of collaboration tie establishment. This average decrease is not statistically significant ($p > 0.05$). For researchers affiliated to institutions other than Beninese institutions, the odds of collaboration tie establishment increases on average by 18.9% compared to internationally affiliated researchers. Overall, model 3 estimated the probability of collaboration tie formation at 11.8% for international researchers, 11.3% for national researchers and 14.2% for regional players.

Since none of the models containing endogenous ERGM terms and/or the dyadic variables, attained convergence, we do not present those results in table 6.3.

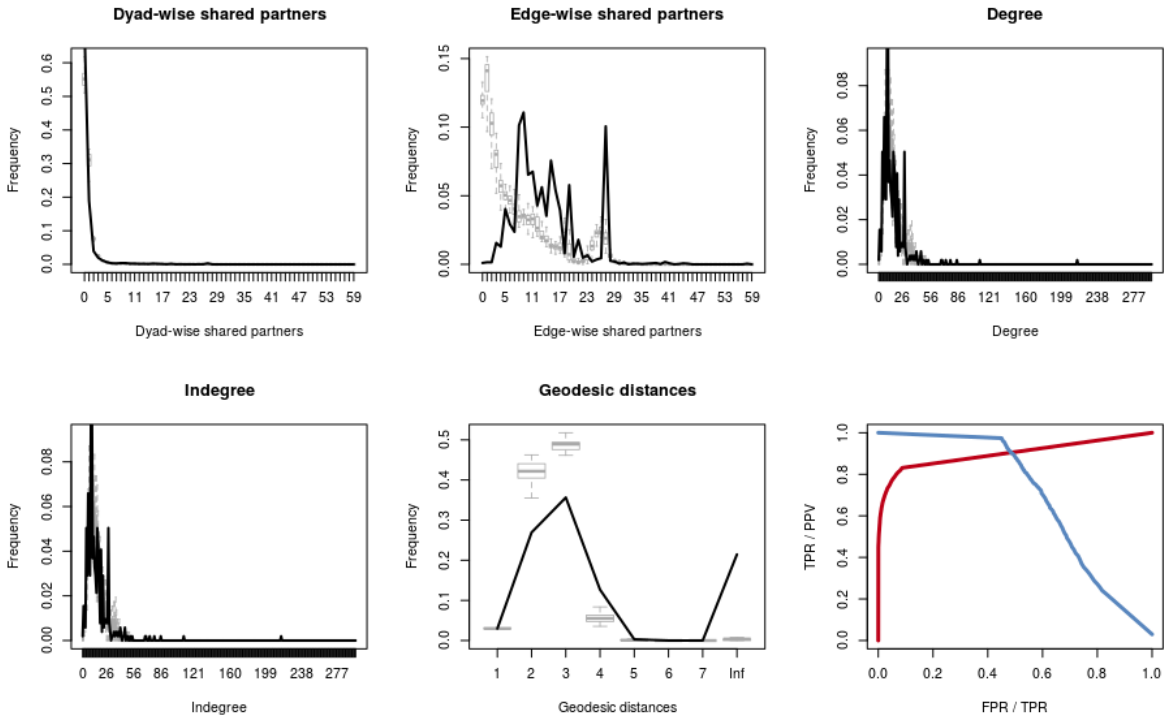


FIGURE 6.10: ERGM goodness-of-fit of final model 3 assessment on the HIV/AIDS co-authorship network.

Figure 6.10 presents the goodness-of-fit of the final model 3. It appears that the ERGM fits well the observed HIV/AIDS co-authorship network in terms of edge-wise, dyad-wise shared partners, degree, geodesic distances, triad census. In addition, 89.9% of the time, model 3 accurately predicted new collaboration ties among the authors ($AUC = 89.9\%$, random models light curves not displayed).

6.3.2.3 Temporal Exponential Random Graph Model

We subset the cumulative observed network in six snapshots according to the following time spans: 1996 – 2001, 2002 – 2008, 2009 – 2010, 2011 – 2012, 2013 – 2014 and 2015 – 2016. In figure 6.11, we show the topological structure of the network snapshots for the different time steps.

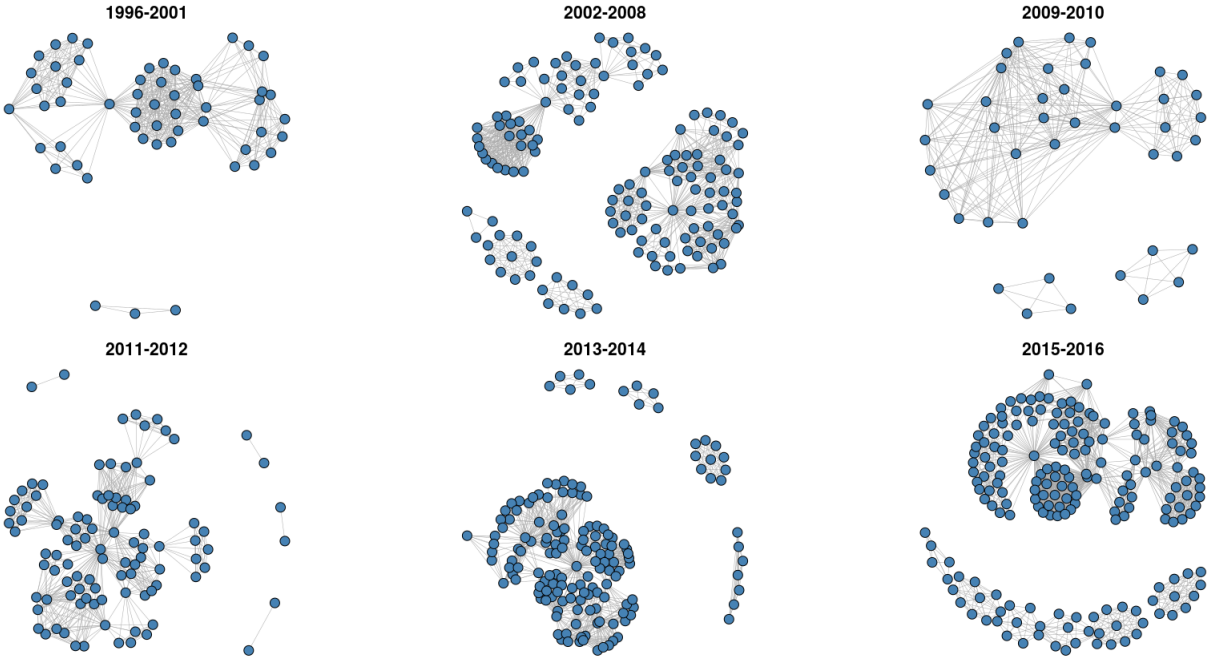


FIGURE 6.11: Topological structure of the different snapshots of the HIV/AIDS co-authorship network.

Table 6.4 summarizes the results of the different temporal models fit to the observed snapshots of the network. The coefficient for the edge term in the null pooled ERGM model 1 is estimated at -5.18 with an associated baseline pooled probability of collaboration tie formation of 0.56%. This probability is lower than the density of the cumulative network. After adjusting for the nodal variables and the homophily terms, model 2 improved slightly over the null model 1. Model 3 adjusted model 2 by including a factor attribute

effect on the collaboration type with a slight improvement over model 2. Model 3 contains the same terms as the final model of the ERGM in the previous section. Unlike the final model of the ERGM, we observed here a significant decrease of 33.6% in the odds of researchers affiliated with Beninese institutions to collaborate compared to international researchers. This effect is maintained after adjusting for the temporal dependencies in model 4.

Model 4 displays a tremendous improvement over model 3, and is hence our final TERGM. The results of model 4 confirm the observation made in section 6.3.2.2 that the process of collaboration tie establishment in the HIV/AIDS network is mainly driven by homophily on collaboration type and on membership to research groups or communities.

Both temporal dependencies effects are significant in the final model. We observed a significantly positive dyadic stability effect accompanied with a significantly negative linear trends effect. For dyadic stability, the coefficient is 0.37 meaning that the odds of existent and non existent collaboration ties at one time point to remain the same at the next time point increased on average by 30.9%. In other words, the odds of new collaboration ties and non-ties to occur from one time point to another is 69.1%. Overall, the probability of international authors to establish a stable collaboration tie is 7.94% versus 6.30% and 9.62% respectively for national and regional researchers.

The goodness-of-fit assessment of the final TERGM model 4 is presented in figure 6.12. The first five subfigures comparing the distribution of endogenous network statistics between the observed network and the simulated ones show a good fit of the final model to the observed network data. The AUC of the ROC curve in the six subfigures is 79.9%

TABLE 6.4. Temporal ERGM of the HIV/AIDS co-authorship network.

	Model 1	Model 2	Model 3	Model 4
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Network structural predictor				
Intercept(edge)	−5.18 (0.02)***	−8.73 (0.05)***	−8.68 (0.06)***	−7.86 (0.09)***
Number of times cited	–	0.00 (0.00)***	0.00 (0.00)***	0.00 (0.00)***
Number of collaborations	–	0.12 (0.00)***	0.11 (0.00)***	0.10 (0.00)***
Number of publications	–	−0.10 (0.01)***	−0.06 (0.01)***	−0.03 (0.01)
Homophily on cluster assignment	–	4.60 (0.05)***	4.61 (0.05)***	4.46 (0.05)***
Homophily on collaboration type	–	0.52 (0.04)***	0.50 (0.04)***	0.59 (0.04)***
Factor attribute effect (collaboration type)				
International	–	–	*REF*	*REF*
National	–	–	−0.29 (0.03)***	−0.25 (0.04)***
Regional	–	–	0.14 (0.03)***	0.21 (0.03)***
Temporal dependencies				
Dyadic stability	–	–	–	0.37 (0.04)***
Linear trends	–	–	–	−0.08 (0.02)***
AIC	5591754.39	5563258.81	5563125.93	3715452.45
BIC	5591781.15	5563352.48	5563246.37	3715595.64
Log Likelihood	−2795875.19	−2781622.40	−2781553.96	−1857715.22

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

for model 4 in predicting ties in the last snapshot. While this performance is lower than the performance of the final ERGM model 3 from the previous section, the walktrap and edge betweenness modularity distributions from model 4 predicted well the observed ones. Finally, the walktrap community comembership prediction displays an AUC of 80%.

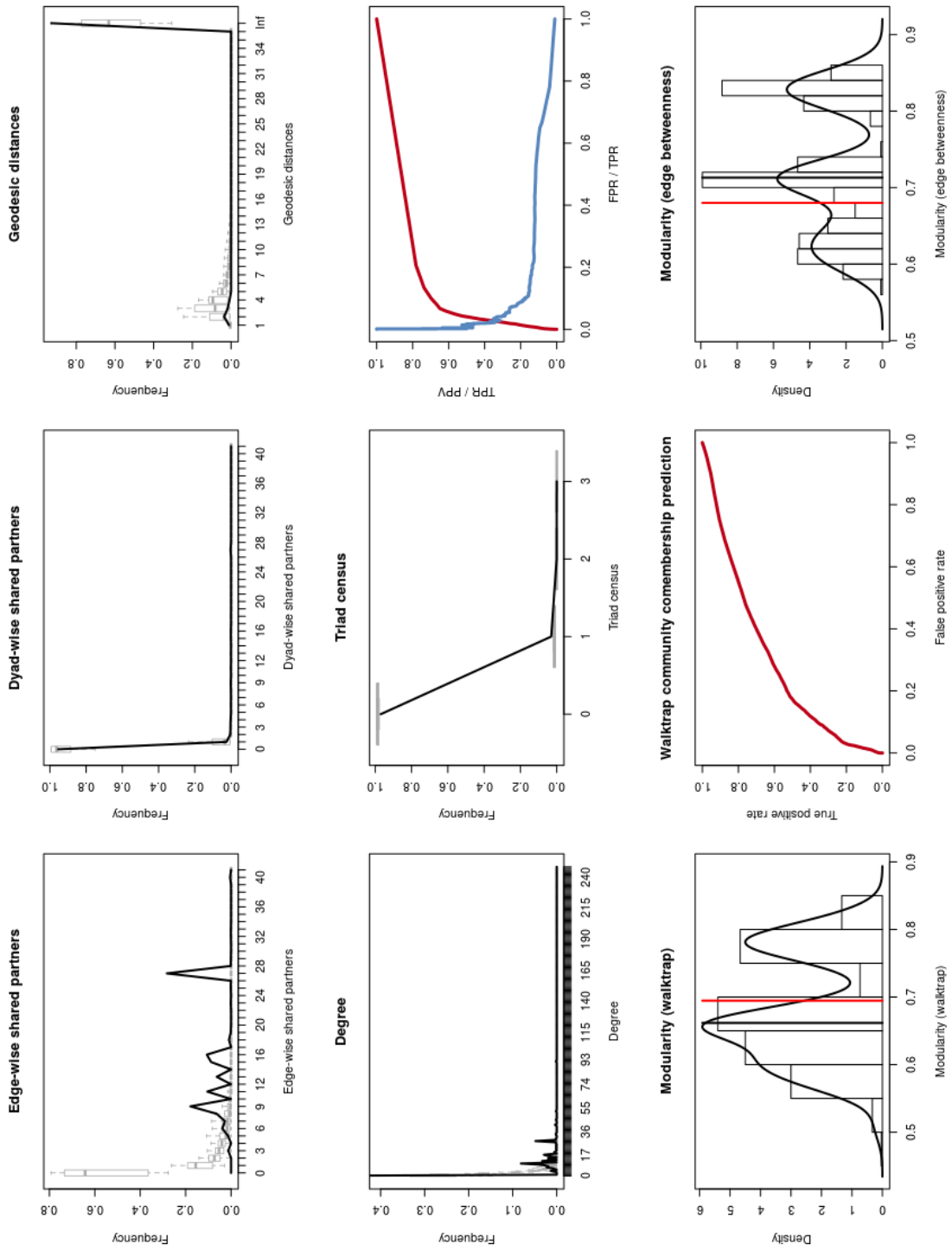


FIGURE 6.12: Goodness-of-fit assessment for the final HIV/AIDS TERGM Model 4 with temporal dependencies of the HIV/AIDS co-authorship network.

6.3.2.4 Latent Network Model

In figure 6.13, we present the 3-dimensional visualization of the HIV/AIDS co-authorship with layouts determined according to the inferred latent eigenvectors from the no pair-specific model (on top), the model containing nodal covariates (middle), and the model containing nodal and dyadic covariates (bottom). Blue vertices represent authors affiliated to Beninese research institutions, Red vertices are authors affiliated to international institutions, Gold vertices represent authors affiliated to African research institutions other than Benin, and White vertices represent authors with no determined affiliations. Vertex sizes are set to be proportional to the betweenness value of each vertex, with bigger vertices emphasizing key broker authors in the network.

The first visualization represents the LNM with no pair-specific covariates. It shows mainly two clusters with little demarcation. We can see that there is a heterogeneity in the spatial distribution of the vertices. After adjusting for the nodal covariates (second visualization), the clustering of the nodes appears less apparent. This results seem to suggest the non-significant role of geography in the establishment of collaboration ties in the HIV/AIDS co-authorship network.

After adding dyadic variables to the model, the resulting visualization shows that there is less structure left to be captured by the latent variables (bottom subfigure on figure 6.13). This observation can explain the failure of our ERGM and TERGM containing dyadic covariates to converge. It also confirms our ERGM and TERGM findings.

We assess the goodness-of-fit of the LNMs. The ROC curves on figure 6.14 shows that the

LNM model containing the nodal covariates ($AUC = 0.966$) outperforms the null model ($AUC = 0.898$).

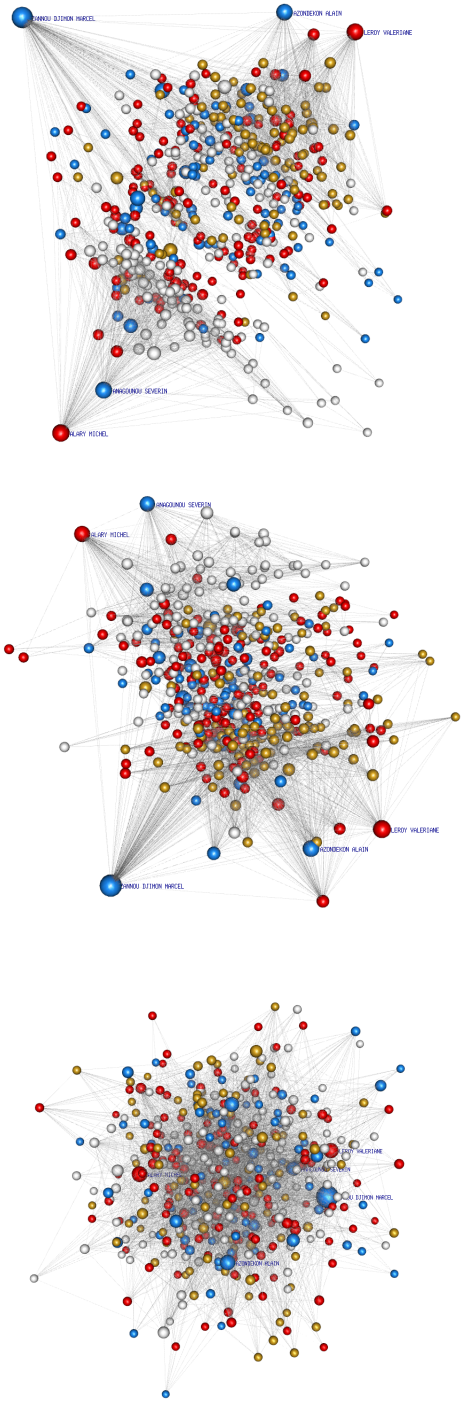


FIGURE 6.13: Visualizations of the HIV/AIDS co-authorship network with layouts determined according to the inferred latent eigenvectors in the LNM models (International (Red); Regional (Gold); Local (Blue); Unknown (White)).

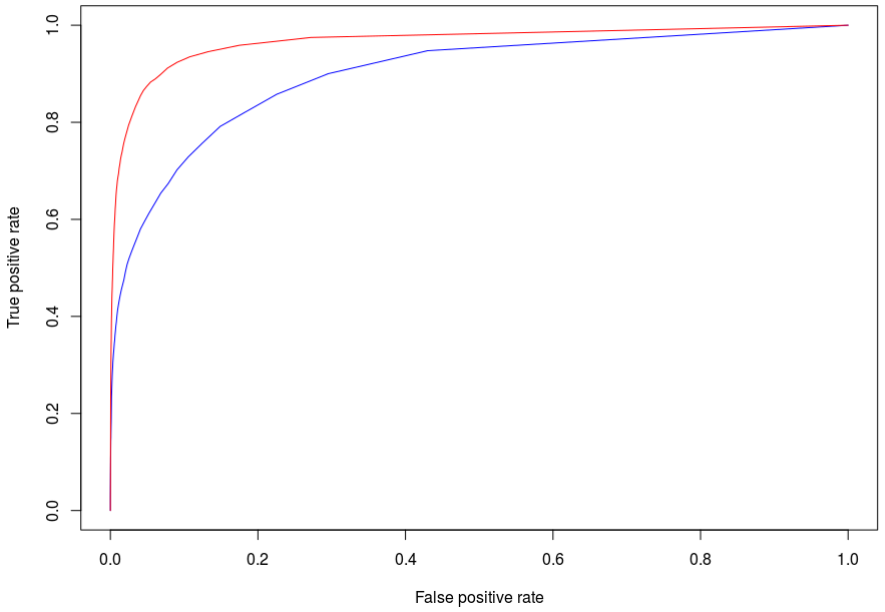


FIGURE 6.14: ROC curves comparing the goodness-of fit of the HIV/AIDS co-authorship network for the model specifying (i) no pair specific covariates (blue) and the model specifying (ii) nodal covariates (red).

6.4 Discussion and Conclusion

This chapter deciphers the HIV/AIDS co-authorship network over the last 20 years. The results from the descriptive analyses in this chapter are similar to the descriptive analyses results from chapter 6. Similar to our findings for the malaria co-authorship network, the HIV/AIDS co-authorship network in Benin is a complex network, as it exhibits unexpected properties that are more extreme than the 4 mathematical models used for the Monte-Carlo based simulations. The observed characteristics disproved previous studies supporting the idea that co-authorship have small world properties [26] or are preferential attachment networks [132]. In fact, unlike our methodology, those studies mainly used descriptive methods and did not apply advance statistical methods to test their network properties. The HIV/AIDS co-authorship network in Benin has a low density with a highly right-skewed node degree distribution. Compared to the malaria co-authorship network, the relatively low transitivity provides evidence of less hierarchy - well connected authors in this network tend to connect with poorly connected ones. This also indicates that this network is less assortative than the malaria co-authorship network, with prolific and non tenure authors connected to similar authors. As in Salamati and Soheili [70], The flow of information in the HIV/AIDS network in Benin is slow as it only relies on 8 authors representing less than 1% of all the authors in the network. The removal of these authors from the network would lead to its collapse. Such a structural vulnerability is not just inherent to the HIV/AIDS co-authorship network, as it is a global observation already reported elsewhere [67]. Since the mathematical models applied here, fell short to

thoroughly explain the mechanistic phenomenon explaining the growth and the structure of the network, we suspect hidden factors which we attempted to model using advanced statistical models.

As our first modeling approach, the SBM identified heterogeneous classes with no dominance of regional, national or international players, despite a reported higher likelihood of Sub-Saharan African countries to collaborate with non-African states [133].

Based on the results from our ERGM and TERGM models, in the HIV/AIDS co-authorship network, authors are more likely to establish collaboration ties within their research groups or communities. Unfortunately, we were not able to control for transitivity as all the models adjusting for this term failed to converge. We suspect the size and complexity of this network to have prevented the convergence of such models, even after 1,000 iterations [134].

Factors such as number of publications, number of citations and number of collaborations were found to have a small but significant ($p<0.001$) association with co-authorship, confirming therefore our first hypothesis. Adding temporal dependencies to our ERGM tremendously improved the fitness of the model to the observed network data, but at a cost of decreased performance compared to the model without temporal dependencies.

The LNM complements the ERGM and TERGM by adding another layer of analysis. With the LNM, we are able to visualize the effect of geography on the structure of the network. The lack of clear cluster demarcation suggests that distance does not play a significant role in collaboration tie formation in the HIV/AIDS network.

Our results confirm that the regain in HIV/AIDS research funding has led to a significant increase in publications number and research collaborations in Benin. In order to consolidate the knowledge generated, there is an urgent need to reinforce the HIV/AIDS research network in Benin given its vulnerability. Identified key brokers and most productive authors need to continuously be supported, and identified junior scientists in the field be promoted.

Chapter 7

Results: The Tuberculosis Co-authorship Network

7.1 Data

The literature search was conducted in the Web Of Science using combinations of TB related MeSH terms including ”Tuberculosis”, ”Mycobacterium”, ”Infection”. The final query set (Table 7.1) returned 109 records. The records were manually screened to verify the involvement of either an author from Benin or the use of data collected in Benin. Overall, 37 documents met the selection criteria. On average, there were 9.38 authors per published document.

After the Author Name Disambiguation, we identified 173 unique authors with a precision of 99.99% and a recall of 99.99%. The generated multigraph co-authorship network

TABLE 7.1. TB Bibliographic Search Queries.

Set	Queries	Results
#1	TOPIC: (Tuberculosis) AND ADDRESS: (BENIN)	109
#2	TOPIC: (Tuberculosis), Refined by: COUNTRIES/TERRITORIES: (BENIN)	77
#3	TOPIC: (Mycobacterium Tuberculosis), AND ADDRESS: (Benin)	77
#4	TOPIC: (Tuberculosis) OR TOPIC: (Infection) AND ADDRESS: (BENIN), Refined by: COUNTRIES/TERRITORIES: (BENIN)	89
Final Set	#1 OR #2 OR #3 OR #4	109

therefore contained 173 vertices (authors) and 1,937 parallel edges (collaborations). As displayed in figure 7.1, we can see the significant increase in publications, scientific collaborations and the number of authors involved in TB research from 2008 until 2016. This general upward trend seems to be linear from the year 2008 to 2016.

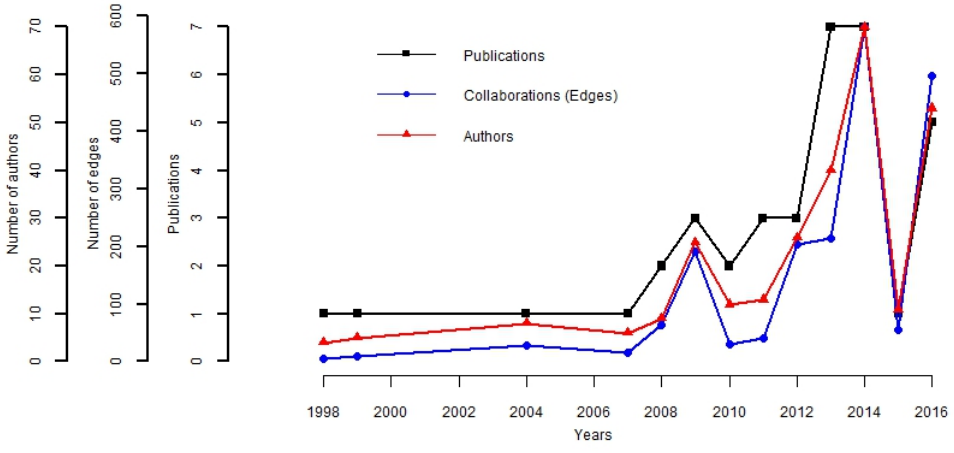


FIGURE 7.1: Evolution of the published TB related documents, authors and collaborations from January 1996 to December 2016

7.2 Descriptive Data Analysis

For the multigraph network, the degree distribution ranged between 2 and 165 with an average degree distribution of 17.36 and a median of 15. In addition, there was a substantial number of vertices with low degrees (Fig. 7.2). The log scale distribution of the degrees on figure 7.3 reveals that there was a tendency of prolific authors to collaborate with less prolific authors.

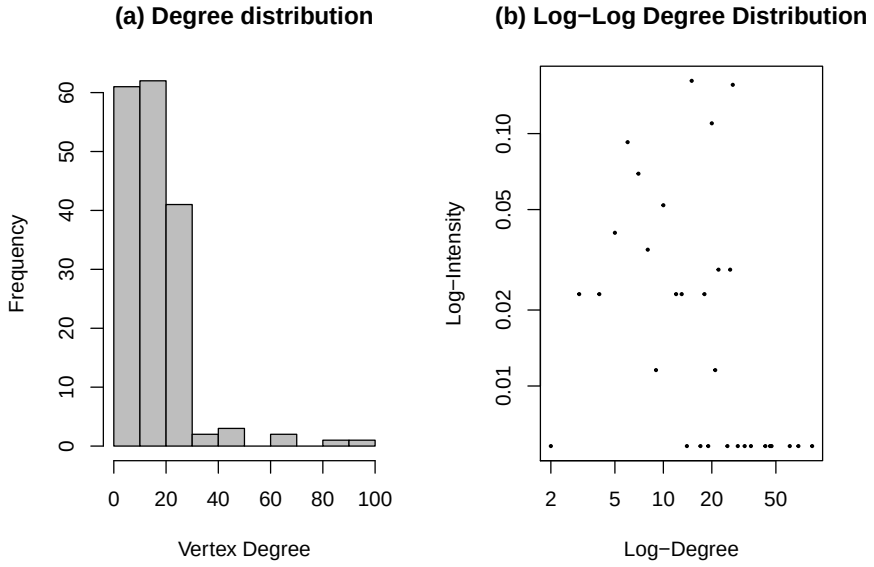


FIGURE 7.2: Degree distribution of the TB co-authorship network

After converting the multigraph network in a weighted graph, the network results in a simple graph of 173 vertices and 1,502 weighted edges. Closeness centrality measures range between 3.68×10^{-5} and 3.28×10^{-4} with a median of 3.18×10^{-4}. Betweenness measures range between 0 and 3,077 with a median of 12.49. A network visualization with the vertices' size proportional to betweenness centrality measures clearly reveals the presence of broker authors (Figure 7.4 and Table 7.2). The median Eigenvectors is 0.087

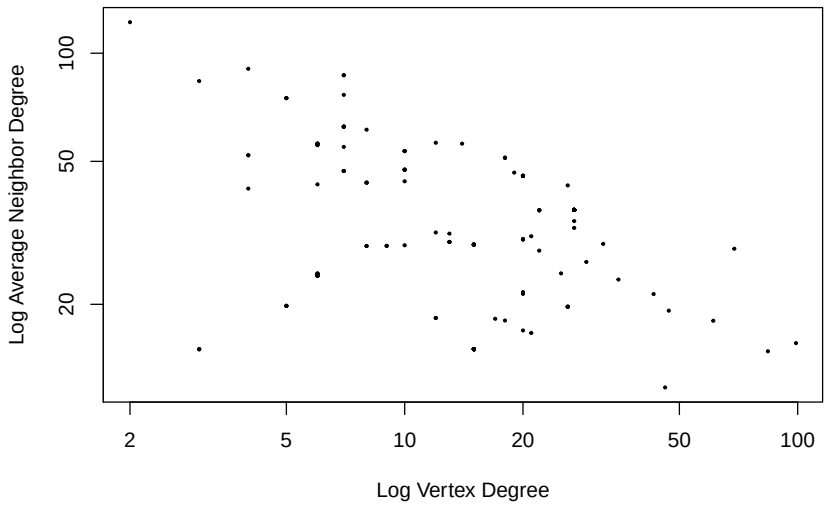


FIGURE 7.3: Log-Average Neighbor degree Distribution of the TB co-authorship network

and a mean of 0.138. Eigenvectors measures reveal the presence of author hubs in the network suggesting the presence of closed collaboration groups. Table 7.2 presents a list of the top 10 author hubs with the highest Eigenvectors values.

Edge betweenness centrality measures identify co-authorship collaboration ties that are important for the flow of information. Table 7.2 presents the top 10 most important collaboration ties for the flow of information in the TB co-authorship network in Benin.

7.2.1 Network Cohesion

Overall, 28 maximal cliques were detected in the network among which 1 clique of size 10, 2 cliques of size 5, and 2 cliques of size 4. The largest clique has size 10.

The TB co-authorship network has a density of 0.10095 indicating that the baseline

TABLE 7.2. List of the most important authors and collaborations in the Tuberculosis co-authorship network

Top 10 Brokers
AFFOLABI DISSOU
GNINAFON MARTIN
DE JONG BOUKE C
TREBUCQ ARNAUD
ODOUN MATHIEU
ANAGONOU SEVERIN
WACHINOU PRUDENCE
FAIHUN FRANK
KASSA FERDINAND
ADE SERGE
Top 10 most connected authors (Top 10 network hubs)
GNINAFON MARTIN
AFFOLABI DISSOU
ANAGONOU SEVERIN
MERLE CORINNE S C
TREBUCQ ARNAUD
OLLIARO PIERO L
RUSTOMJEE ROXANA
LO MAME BOCAR
LIENHARDT CHRISTIAN
HORTON JOHN
Top 10 most important edges for information flow
ODOUN MATHIEU – GNINAFON MARTIN
FAIHUN FRANK – DE JONG BOUKE C
ODOUN MATHIEU – TREBUCQ ARNAUD
ZELLWEGER J P – GNINAFON MARTIN
TREBUCQ ARNAUD – ADJONOU CHRISTINE
ODOUN MATHIEU – WACHINOU PRUDENCE
AFFOLABI DISSOU – BAHSOW OUMOU
AFFOLABI DISSOU – TOUNDOH N
AFFOLABI DISSOU – BEKOU W
AFFOLABI DISSOU – MAKPENON A
Weak articulation point
WACHINOU PRUDENCE

probability of collaboration tie formation is 10.095%. The network also has a transitivity of 0.6305 meaning that 63.05% of the connected triples in the network are close to form triangles. The transitivity measures the global clustering of the network.

The network is not connected and a census of all the connected components within the

network reveals the existence of a giant component that dominates all the other connected components. The giant component of the TB co-authorship network includes 90.8% (157 vertices) of all the vertices in the network with the other components alone carrying less than 0.1% of the vertices in the network (Fig. 7.4).

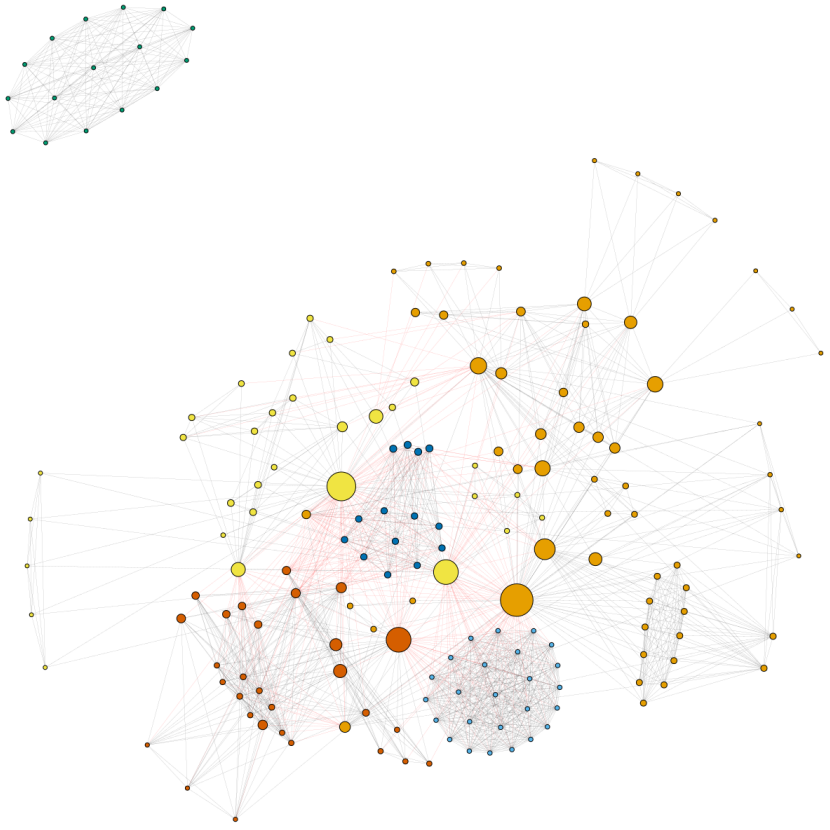


FIGURE 7.4: Topological Structure of the Tuberculosis co-authorship network. Authors (vertices) of the same color belong to the same research community or cluster

Information flow assessment of the network via cut vertices reveals the existence of a single author as the most vulnerable vertex in the network (Table 7.2). The cut vertex constitute the weak articulation point of the TB co-authorship network. Cut vertices

represent a measure of the vulnerability of the network [98].

The agglomerative hierarchical clustering method identifies 6 different clusters in the network. Sizes of the clusters range between 14 and 58 authors. Out of the 6 research clusters detected, 5 are in the giant component. Figure 7.4 displays the giant component of the network with each different colors representing each of the 6 clusters.

7.3 Modeling

7.3.1 Mathematical Modeling

From the hierarchical clustering method of community detection, 6 different clusters were detected in the co-authorship network out of which 5 form a giant component. We performed 1,000 Monte Carlo based simulations to test the significance of this observed characteristic of the TB co-authorship network. Figure 7.5 clearly demonstrates that the number of communities detected is unusual from the perspective of both Classical random graphs and generalized random graphs (p-value < 0.0001). From the Classical random graph model, the expected number of communities was 4.734 (95%CI: 4.70 – 4.77). Similarly, the expected number of communities from the generalized random graph model is 5.34 (95%CI: 5.29 – 5.38).

Figure 7.6 displays the number of detected clusters or research communities using the Barabási-Albert's preferential attachment and the Watts-Strogatz models. Here too, the observed number of communities was extreme per both models (p-value < 0.0001). The

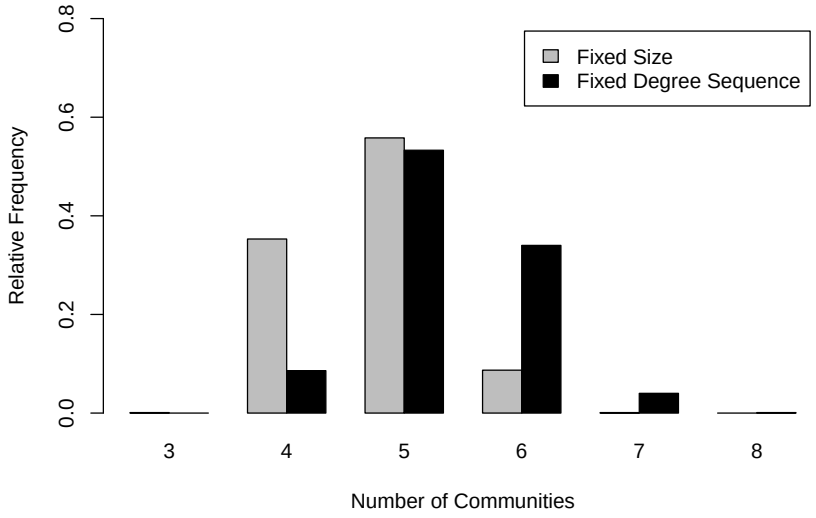


FIGURE 7.5: Monte-Carlo simulations of the TB network: Number of detected communities by the random graph models

expected number from the Watts-Strogatz model simulations is 3.017 (95%CI: 3.01 – 3.03) and 13.77 (95%CI: 13.70 – 13.85) from the Barabási-Albert model simulations.

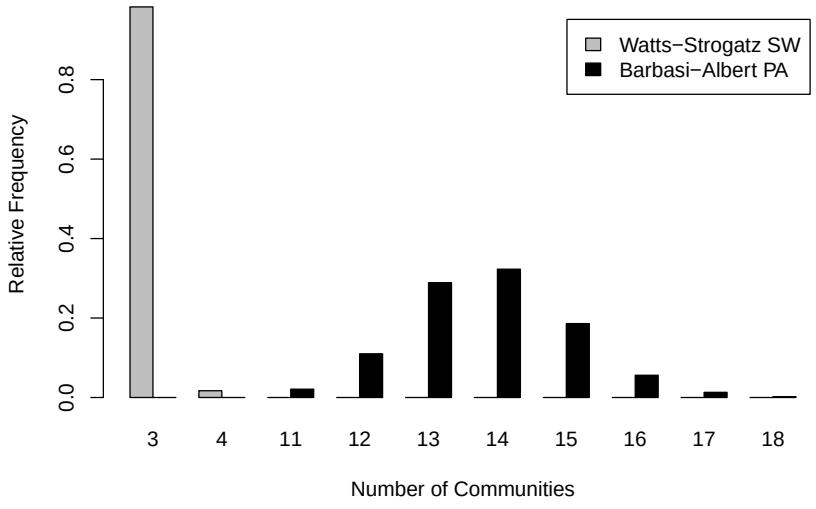


FIGURE 7.6: Monte-Carlo simulations of the TB network: Number of detected communities by the Watts-Strogatz and the Barabási-Albert models

We also compared the clustering coefficient and the average shortest-path length. Let's recall that the observed clustering coefficient is 0.614. On one hand, there was substantially more clustering in our TB co-authorship network than expected from both random graph models (p-value < 0.0001). The expected clustering coefficients was 0.10087 (95%CI: 0.10068 – 0.10107) and 0.1937 (95%CI: 0.1934 – 0.1939) respectively for the classic random graph and the generalized random graph models.

On the other hand, there was substantially less clustering in our TB co-authorship network than expected from the Watts-Strogatz Small World model which expected clustering was 0.7259 (95%CI: 0.7258 – 0.7260).

We observed an average shortest-path length of 2.126 in the TB co-authorship network. This observed shortest-path length is significantly larger than what was expected from the random graph models (p-value < 0.0001) and significantly lower than what was expected from Watts-Strogatz small world model and the Barabási-Albert preferential attachment model (p-value < 0.0001).

The average shortest-path length was 2.0548 (95%CI: 2.0546 – 2.0550) and 2.072 (95%CI: 2.0715 – 2.0726) respectively for the classic random graph and the generalized random graph models.

For the Watts-Strogatz small world and the Barabási-Albert models, the average shortest-path length is respectively 2.623 (95%CI: 2.616 – 2.631) and 6.06 (95%CI: 6.03 – 6.09).

We performed the same simulations on the giant component of the network with similar results leading to the same outcomes.

7.3.2 Statistical Modeling

7.3.2.1 Stochastic Block Model

The SBM identifies 14 classes with a degree of latitude of 9 to 14 classes being reasonable (See ICL plot on figure 7.7).

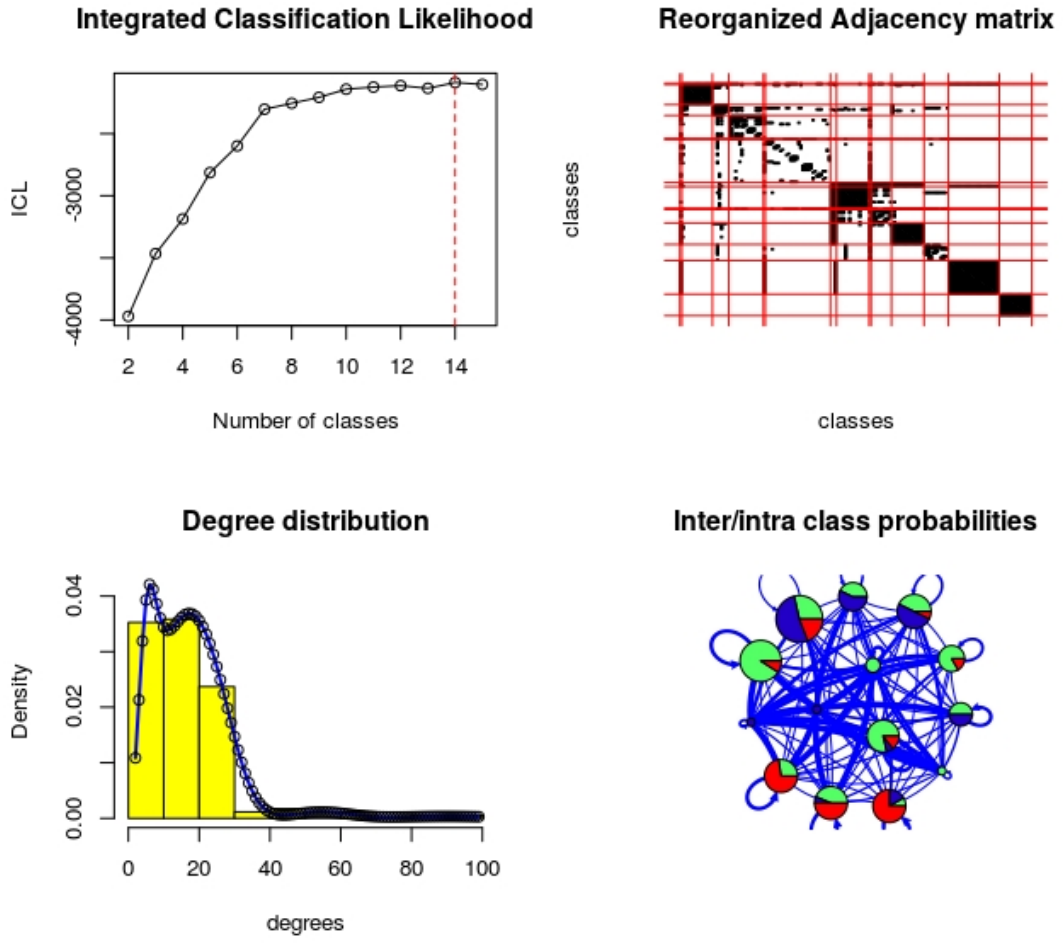


FIGURE 7.7: Summary of the goodness-of-fit of the SBM analysis on the Tuberculosis co-authorship network.

The fitted SBM describes well the observed degree distribution. The vertices in the network depicting the inter/extra probabilities represent the 14 identified classes, with each

one of them divided into a pie chart displaying the proportion of authors of international affiliations (lightgreen), authors of regional or other African affiliations (red), and authors affiliated to Beninese research institutions (blue). Generally, the dominance across the classes of international and regional players is observed. From the inter/intra probability network shows denser inter class ties. Looking at the pie charts, we can see that the classes are heterogeneous with most of the classes having the same sizes (7.7). Figure 7.8 presents the distribution of the classes by affiliation types.

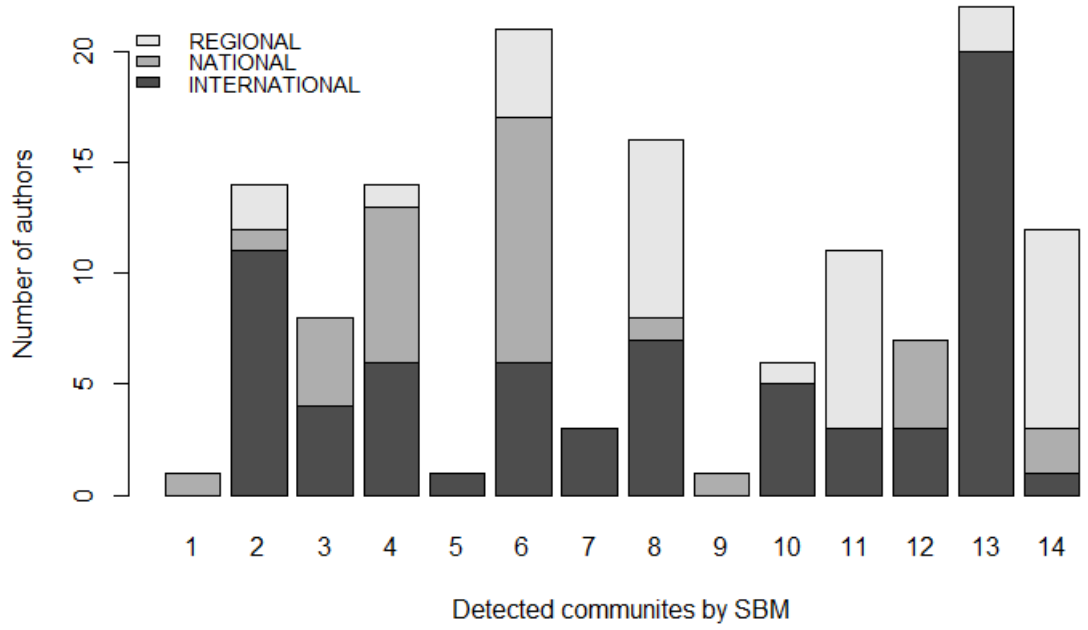


FIGURE 7.8: Distribution of national, international and regional authors by communities detected by the SBM in the TB network.

	Model 1	Model 2	Model 3
	Estimate (SE)	Estimate (SE)	Estimate (SE)
Network structural predictor			
Intercept(edge)	-2.19 (0.03)***	-7.84 (0.16)***	-7.86 (0.17)***
Number of times cited	–	0.01 (0.00)***	0.01 (0.00)***
Number of collaborations	–	0.08 (0.00)***	0.07 (0.00)***
Number of publications	–	-0.05 (0.01)**	0.01 (0.02)
Homophily on cluster assignment	–	6.02 (0.13)***	6.12 (0.14)***
Homophily on collaboration type	–	0.83 (0.10)***	0.90 (0.10)***
Factor attribute effect (collaboration type)			
International	–	–	REF
National	–	–	-0.40 (0.09)***
Regional	–	–	0.22 (0.08)**
AIC	9737.42	3776.48	3747.34
BIC	9745.03	3822.12	3808.20
Log Likelihood	-4867.71	-1882.24	-1865.67

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

TABLE 7.3. ERGM of the TB co-authorship network.

7.3.2.2 Exponential Random Graph Model

We fit multiple ERGMs (Table 7.3). In the null model (model 1), the inverse logit of the coefficient associated with the intercept (edge term) is 0.10 which is the baseline probability of collaboration tie establishment and also the density of the TB co-authorship network.

Model 2 including all nodal variables, a homophily term on collaboration type and on cluster assignment improved tremendously compared to model 1 (See AIC, BIC and model likelihood in table 7.3). We note a decrease in the edge effect (Coefficient $= -7.84$, $p < 0.001$) with the associated conditional probability (given all the other terms in the model) estimated at 0.039%. For the remaining terms in model 2, we observed a positive

and significant effect except for the number of publications. Model 3 including the collaboration type as factor term, improved substantially compared to model 2. We therefore chose model 3 as our final model. One unit increases the number of citation, increases the odds of collaboration ties establishment by 1%. A one unit increase in the number of collaborations is associated with a 7.25% increase in the odds of collaboration ties establishment. The coefficient associated with the number of publications is insignificant. Model 3 further proves that the process underlying the structure of the TB co-authorship network in Benin is mainly driven by homophily on cluster assignment or membership to a research community or group (Coefficient = 6.12, $p < 0.001$). The conditional probability of any two authors belonging to the same research group is estimated at 14.93% compared to the baseline probability of 10%. The same probability changes to 30.15% after adjustment by the collaboration type, and 32.08% after adjusting for the number of citations, collaborations and publications. Compared to research affiliated to international institutions, researchers affiliated to Beninese institutions have 49.2% average decrease in the odds of collaboration tie establishment. This average decrease is not statistically significant ($p > 0.05$). For researchers affiliated to institutions other than Beninese institutions, the odds of collaboration tie establishment increase on average by 24.05% compared to internationally affiliated researchers. Overall, model 3 estimated the probability of collaboration ties formation at 32.08% for international researchers, 24.05% for national researchers and 37.05% for regional players.

Unfortunately, none of the models containing endogenous ERGM terms and/or the dyadic variables, attained convergence, we do not present those results in table 7.3.

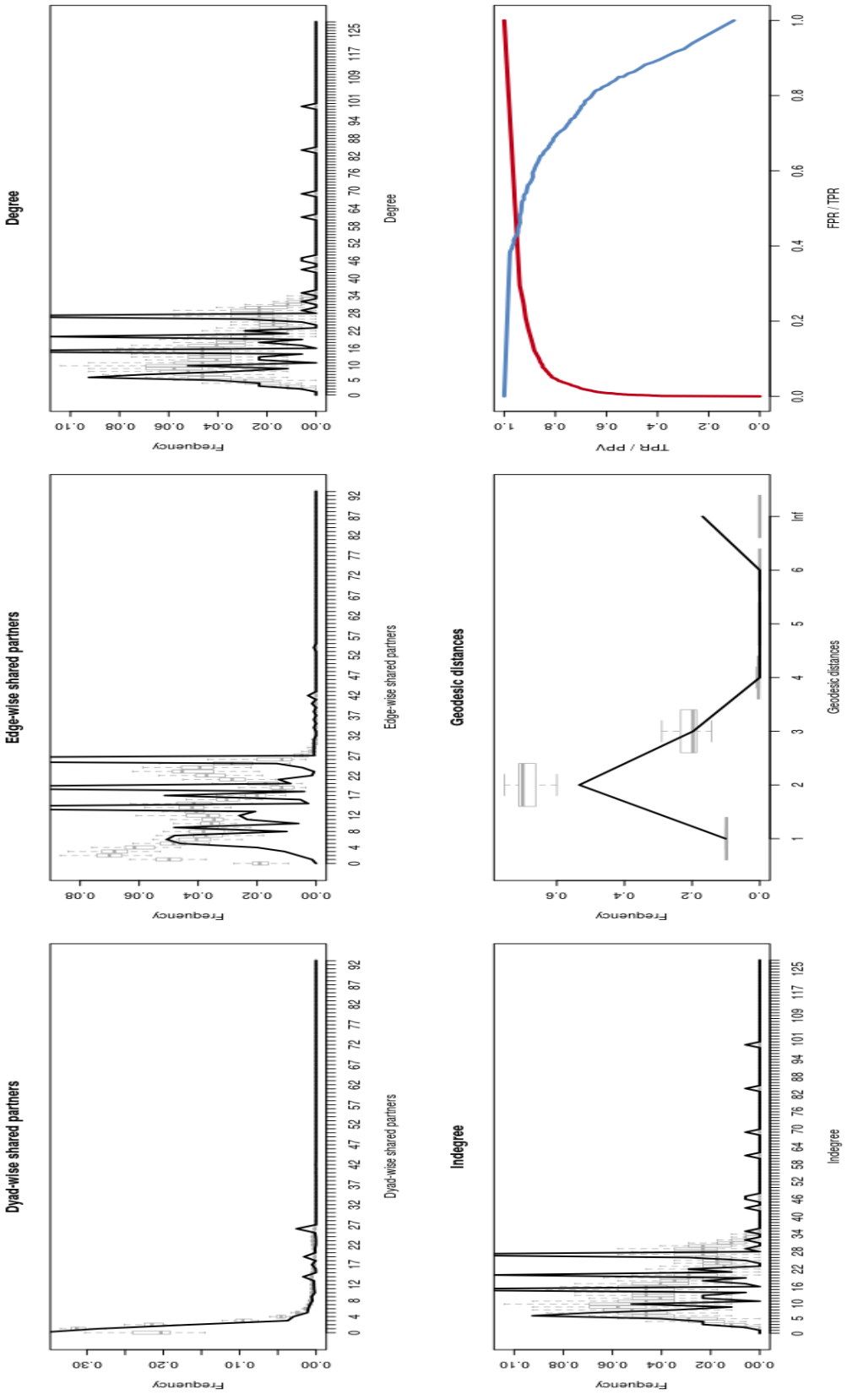


FIGURE 7.9: ERGM goodness-of-fit of final model 3 assessment on the TB co-authorship network.

Figure 7.9 presents the goodness-of-fit of the final model 3. It appears that the ERGM fits somewhat poorly the observed TB co-authorship network in terms of edge-wise, dyad-wise shared partners, degree, geodesic distances, triad census. Meanwhile, it displays a 93.7% for the ROC model (in red) and 80.9% for the Precision Recall (PR) model.

7.3.2.3 Temporal Exponential Random Graph Model

We subset the cumulative observed network in five snapshots according to the following time spans: 1996 – 2008, 2009 – 2011, 2012 – 2013, 2014 – 2015 and 2016. In figure 7.10, we show the topological structure of the network snapshots for the different time steps.

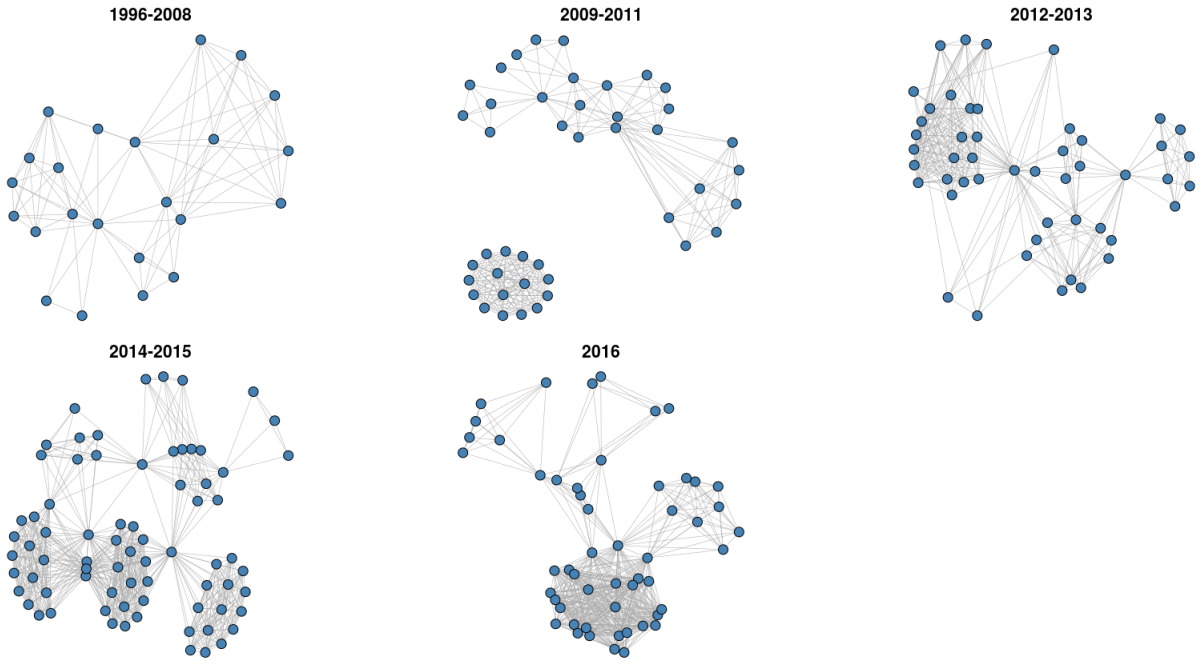


FIGURE 7.10: Topological structure of the different snapshots of the TB co-authorship network.

Table 7.4 summarizes the results of the different temporal models fit to the observed snapshots of the network. The coefficient for the edge term in the null pooled ERGM model

1 is estimated at -3.75 with an associated baseline pooled probability of collaboration tie formation of 2.30%, which is lower than the density of the observed cumulative TB network.

Model 2 adjusts for the nodal variables and the homophily terms improved slightly over the null model 1. Model 3 adjusted model 2 by including a factor attribute effect on the collaboration type with a slight improvement over model 2. Unlike the final model of the ERGM, we observed in model 3, a significant decrease of 23.4% in the odds of researchers affiliated with Beninese institutions to collaborate compared to international researchers. This percentage decrease changes to 40.5% after adjusting for the temporal dependencies in model 4.

We chose Model 4 as our final model because it significantly improved over model 3. The results of model 4 confirm our observation from the ERGM results that the process of collaboration tie establishment in the TB network is mainly driven by homophily on collaboration type and on membership to research groups or communities.

Temporal dependencies effects proved significant in the final model. A significantly positive dyadic stability effect accompanied with a significantly negative linear trends effect is observed. For dyadic stability, the coefficient is 0.44 meaning that the odds of existent and non existent collaboration ties at one time point to remain the same at the next time point increased on average by 35.6%. In other words, the odds of new collaboration ties and non-ties to occur from one time point to another is 64.4%. Overall, the probability of international authors to establish a stable collaboration tie is 15.71% versus 11.71% and 16.11% respectively for national and regional researchers.

TABLE 7.4. Temporal ERGM of the TB co-authorship network.

	Model 1	Model 2	Model 3	Model 4
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Network structural predictor				
Intercept(edge)	−3.75 (0.02)***	−10.07 (0.15)***	−10.01 (0.16)***	−8.62 (0.28)***
Number of times cited	–	0.00 (0.00)*	0.00 (0.00)	−0.00 (0.00)**
Number of collaborations	–	0.14 (0.00)***	0.14 (0.00)***	0.16 (0.00)***
Number of publications	–	0.68 (0.03)***	0.72 (0.03)***	0.57 (0.03)***
Homophily on cluster assignment	–	5.24 (0.11)***	5.23 (0.11)***	5.40 (0.13)***
Homophily on collaboration type	–	0.69 (0.08)***	0.69 (0.08)***	0.73 (0.09)***
Factor attribute effect (collaboration type)				
International	–	–	*REF*	*REF*
National	–	–	−0.21 (0.07)**	−0.34 (0.08)***
Regional	–	–	0.03 (0.07)	0.03 (0.08)
Temporal dependencies				
Dyadic stability	–	–	–	0.44 (0.07)***
Linear trends	–	–	–	−0.36 (0.06)***
AIC	431184.00	419860.54	419853.82	253170.25
BIC	431205.66	419936.36	419951.30	253284.48
Log Likelihood	−215590.00	−209923.27	−209917.91	−126574.12

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The goodness-of-fit assessment of the final TERGM model 4 is presented in figure 7.11. Regarding the endogenous network statistics, we observe a better fit of the final TERGM model 4 compared to the final ERGM model 3. In other words, the simulated network by model 4 show a good fit to the observed TB network data. The AUC of the ROC curve of model 4 (see dark red curve on subfigure 6) is estimated at 83.2% meaning that 83.2% of the times, model 4 accurately predicts ties in the last snapshot. While this performance is lower than the performance of the final ERGM model 3 from the previous section, the walktrap and edge betweenness modularity distributions from model 4 predicted well the observed ones. Finally, the walktrap community comembership prediction displays an AUC of 71.4% (see dark red curve on subfigure 5).

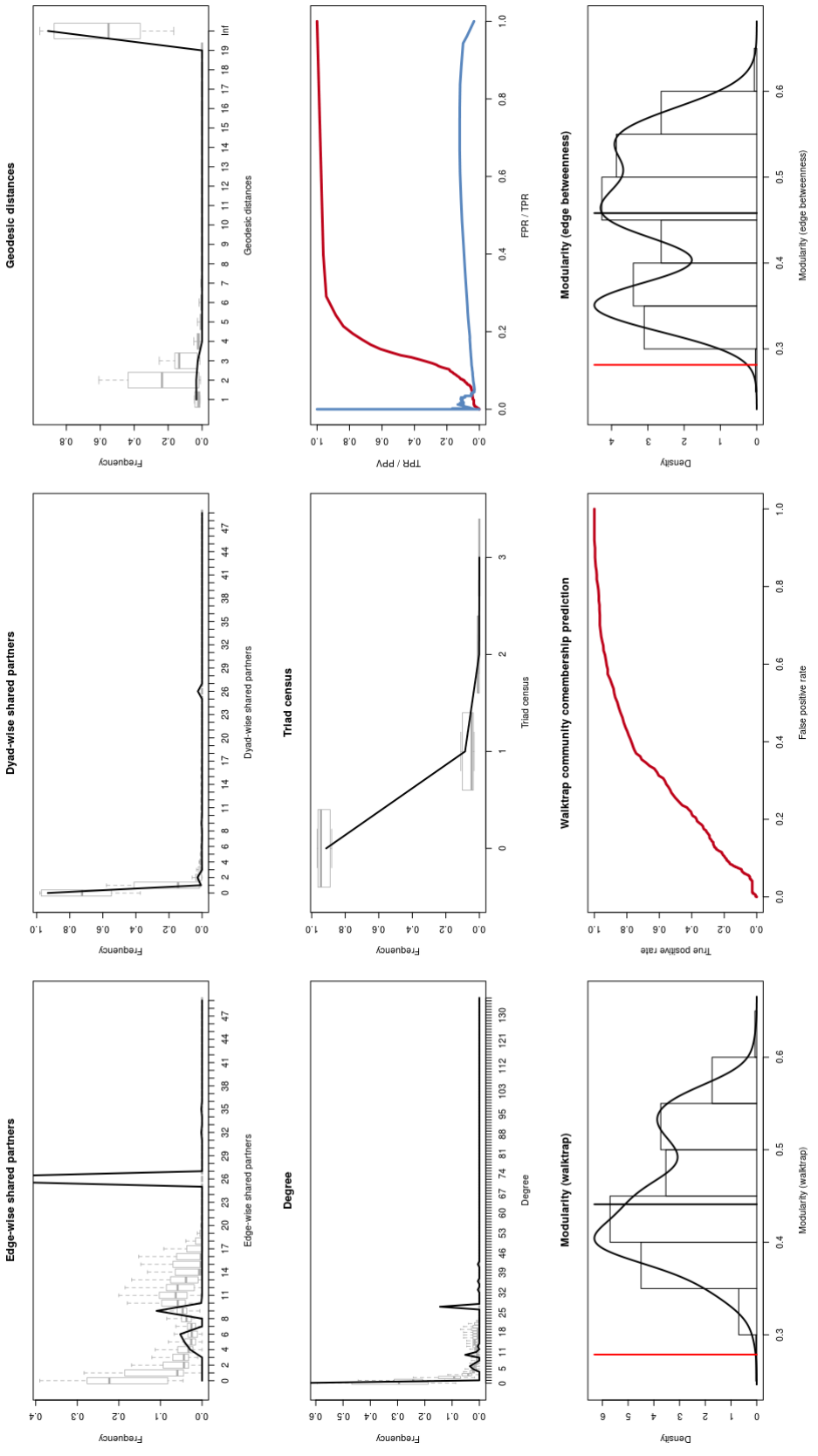


FIGURE 7.11: Goodness-of-fit assessment for the final TB TERGM Model 4 with temporal dependencies of the TB co-authorship network.

7.3.2.4 Latent Network Model

On the 3-dimensional visualization of the TB co-authorship network presented on figure 7.12, the layouts are determined according to the inferred latent eigenvectors from the no pair-specific model (on top), the model containing nodal covariates (middle), and the model containing nodal and dyadic covariates (bottom). Blue vertices represent authors affiliated to Beninese research institutions, Red vertices are authors affiliated to international institutions, Gold vertices represent authors affiliated to African research institutions other than Benin, and White vertices represent authors with no determined affiliations. Vertex sizes are set to be proportional to the betweenness value of each vertex, with bigger vertices emphasizing key broker authors in the network.

The first visualization represents the null LNM with no pair-specific covariates. It shows mainly three clusters. The largest cluster appears more spatially heteregeneous than the other two. It is also the largest cluster that contains the majority of the authors affiliated with Beninese research institutions. The other two clusters seem to be dominated respectively by international and regional researchers. This model fits reasonably well to the observed TB network ($AUC = 0.912$). This observation suggests a significant effect of geography in the odds of collaboration tie establishment. After adjusting for the nodal covariates (second visualization), there is less structure left to be captured by the latent variables and the clustering is no more apparent. Adding dyadic attributes to the model leads to similar outcome despite an increase in terms of performance ($AUC = 0.974$).

On figure 7.13, we present the ROC curves of each of the LNM models containing the nodal covariates and the null model.

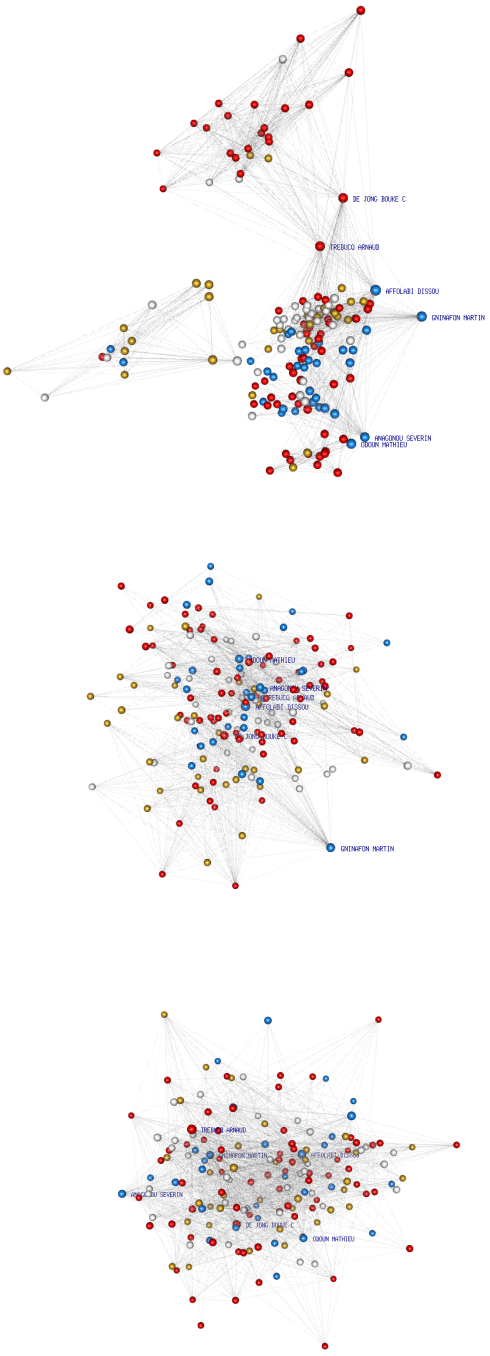


FIGURE 7.12: Visualizations of the TB co-authorship network with layouts determined according to the inferred latent eigenvectors in the LNM models (International (Red); Regional (Gold); Local (Blue); Unknown (White)).

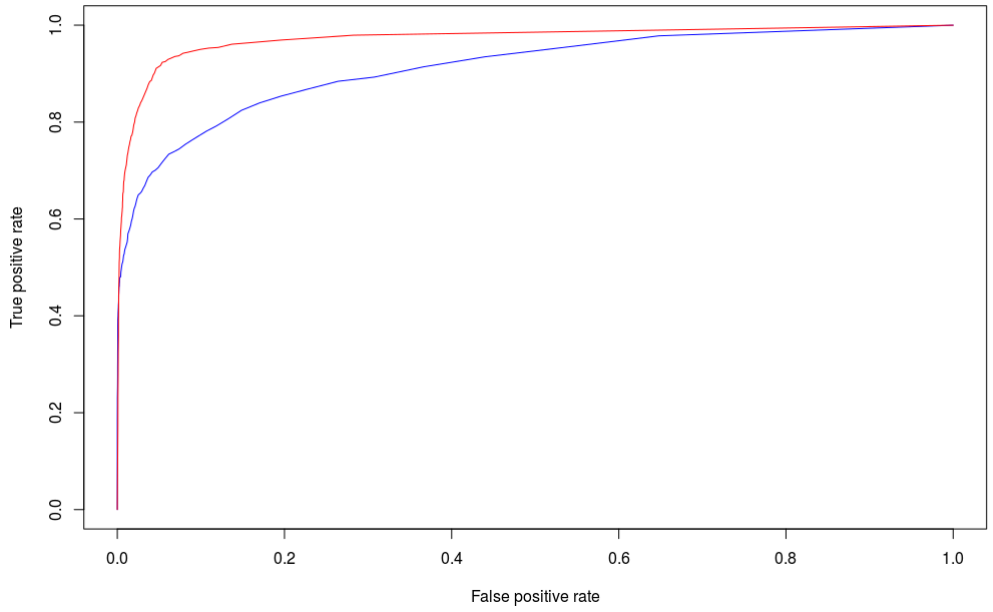


FIGURE 7.13: ROC curves comparing the goodness-of fit of the TB co-authorship network for the model specifying (i) no pair specific covariates (blue) and the model specifying (ii) nodal covariates (red).

7.4 Discussion and Conclusion

This chapter provides insights in the structural characteristics of the TB co-authorship network in Benin over the last 20 years. The evolution of the number of publications, authors and collaboration ties suggests a linear growth over the investigation period. We expected such findings given the place of TB in the public health concerns of Benin and the intensive effort towards the reduction of the incidence and the numerous campaigns of sensibilization [141]. The findings from the descriptive analysis suggest that the mechanism underlying the formation of the TB co-authorship network in Benin is not random. However, we found inconclusive evidence of small world properties that further Monte-Carlo simulations disproved. The presence of closed research groups is supected given the non-trivial number of authors with higher order of magnitudes. The observed trend of prolific authors in the TB network to collaborate with less prolific ones is another indication suggesting that TB research is a low productivity research field in Benin. Only 37 published documents were found relevant to the present study. In fact, none of the top 10 key brokers in our TB co-authorship network, was on the list of the top most connected authors and therefore would suggest the relative absence of long publishing tenure authors in the network [129].

The flow of information in the TB co-authorship network in Benin is slow as it only relies on a single author. A study by Salamatia and Soheili [70] on a co-authorship analysis of Iranian researchers in the field of violence reported similar but less extreme findings. For Bales et al. [130, 131], the most important authors in co-authorship networks generally

tend to be the ones with the highest degree of collaborations. For information flow, cut vertices provide a better approach to identifying vertices that are important to the long-term substainability of co-authorship networks [98]. The only author identified as a cut vertex is therefore the most important author for information flow.

Our observed network has unexpected properties compared to classic small-world networks. Our TB co-authorship network displays properties that are more extreme than those of small-world and preferential attachement networks contradicting previous studies reporting co-authorship network as having small-world or preferential attachment properties [26, 132].

As the first advanced statistical model we applied to this network, the SBM identified heterogeneous classes with higher probabilities towards inter class ties establishment. This observation is different from what we observed for the malaria and the HIV/AIDS co-authorship network which both display low inter class probabilities and higher intra class probabilities of tie formation.

As in the malaria and the co-authorship network, the ERGM and TERGM results suggest that authors within the TB co-authorship network are more likely to establish collaboration ties within their research groups or communities. Although marginal, factors such as number of publications, number of citations and number of collaborations are associated to higher likelihood to establishing collaboration ties, confirming therefore our first hypothesis. Adding temporal dependencies to our ERGM models tremendously improved the fitness of the model to the observed network data, but at a cost of decreased performance compared to the model without temporal dependencies.

We expected the ERGMs and TERGMs containing ERGM structural terms to converge for the TB co-authorship network given its relatively smaller size. Unfortunately, as for the malaria and the HIV/AIDS co-authorship networks, adding such terms to the models proved computationally expensive. None of the models converged after $1,000$ iterations. We therefore, suspect the complexity of the network to have prevented the convergence of the models containing structural ERGM terms [134].

With the LNM, we complement the ERGM and TERGM by adding an extra layer of analysis. Visualizing the effect of geography on the structure of the network, we notice that none of the nodal or dyadic covariates played a significant role in the spatial distribution of the network. Such an observation contradicts that of the HIV/AIDS co-authorship network. The cluster demarcation observed with the null LNM suggests that distance does play a significant role in collaboration tie formation in the TB co-authorship network.

As the co-infection TB-HIV/AIDS continues to be an important aspect of the public health strategies in the Republic of Benin, consolidating the knowledge generated from the TB-related research is crucial. Furthermore, public health policies must empower and reinforce the different research groups or communities involved in the research effort. Our results suggest a need for a continuous support to the TB research network, considering its low productivity status in Benin. Such actions will help stabilize the research groups already involved in TB research and promote the junior scientists in the field. We finally believe that such measures will ultimately insure the long-term sustainability of the TB co-authorship and collaborative research network in Benin.

Chapter 8

AuthorVis: A Co-authorship Visualization and Scientific Collaboration Prediction tool

8.1 Background

In this chapter, we describe a co-authorship network exploration, and link prediction tool we created and that is specific to the three networks investigated in this dissertation. While many network visualization solutions have already been proposed, most of them are not specifically adapted to co-authorship networks [78, 80, 83, 142]. Even those designed for visualizing co-authorship networks have several limitations among others, their inability to satisfactorily display large networks, the lack of interactivity in the display,

and the inability for the end user to control the display [83].

Here, we present a tool that not only addresses those limitations, but provides a visualization of each of the networks and allows the end user to query each network. Our approach integrates bibliometrics information to the visualization. In our design model, all the authorship information are embedded within the display of the network. In the visualization interface, users can select a particular node or author to emphasize its subnetwork, hover over a node to display author's information or select an edge between two vertices/authors to display information related to materials co-authored by the two vertices defining that particular edge.

8.2 Data

Currently, **AuthorVis** is designed specifically for the visualization of the Malaria, Tuberculosis and HIV/AIDS collaborative network in Benin. We refer the reader to section 4.2 for details on the collection and treatment of the co-authorship data. On the server end, each network data is maintained as an igraph object. Each submitted user query is interpreted and incorporated in an igraph function to extract the network data. Another igraph object is generated as a result and converted into a JSON data using an executable Python script that we provided within the tool.

8.3 Programmer View

8.3.1 Design and Architecture

AuthorVis is implemented as a Shiny dashboard with an R based backend system that manages each co-authorship network data as an igraph object [143]. The backend server side is a combination of a Shinyserver and an HTTP server (Figure 8.1). The Shiny application is built using the **Shinyboard** [144, 145] R package. A set of R scripts manages global libraries (`global.R`), controls the dashboard user interface (`ui.R`), and handles backend processings (`server.R`). The user interface script (`ui.R`) communicates with the frontend dashboard interface on the client side and the backend processing script (`server.R`) on the Shinyserver. When the user submits a request, it is passed from the dashboard interface to `server.R` via `ui.R`. The request is subsequently processed, and the output is transferred to the dashboard on the client side via `ui.R`. However, when the request is a query to explore and visualize a co-authorship network, the server also outputs a graph object which is converted into JSON graph file thanks to a python script. The graph file is then transferred to the HTTP server to be displayed on the Network Visualization Interface. The HTTP server has been implemented with the Node.js built-in HTTP module. Node.js is a Javascript server-side platform for the development of web servers [146]. The front-end Network Visualization Interface is handled by the HTTP web server which renders the JSON graph object into an HTML file (`index.html`). A script (`code.js`) written in Javascript using the Javascript D3.js [82] library handles user interactivity and the control of the display. D3.js or Data-Driven Documents has

been designed for manipulating documents based on data and to generate interactive and dynamic data visualizations in web browsers (Figure 8.1).

8.4 User View

8.4.1 Shiny Dashboard Interface

The frontend Shiny dashboard interface has five menu options displayed on its left sidebar. The network query and exploration interface is accessible from the ”Explore Network” menu option and the link prediction interface is accessible via the ”Prediction” menu option. Other options in the side bar menu include the ”Codes” menu where we share the Shiny dashboard scripts, the ”Readme” menu displaying a documentation for the tool, and the ”About” menu which provides general information on the tool (See subfigure (d) on figure 8.2).

When the user selects the ”Explore Network” menu option, the dashboard brings him to the appropriate page containing a simple query builder. After selecting a network, the user can define a time period and may search for a specific author or set of authors. As the user builds his query, the dashboard responds interactively, displaying the number of vertices and edges returned by the query. When the user clicks on the ”Query Network!” button, the query is submitted to the server. Once the processing is done, a URL is displayed and the user is prompted to click on it to launch the Network Visualization Interface (subfigures (a) and (b) on figure 8.2).

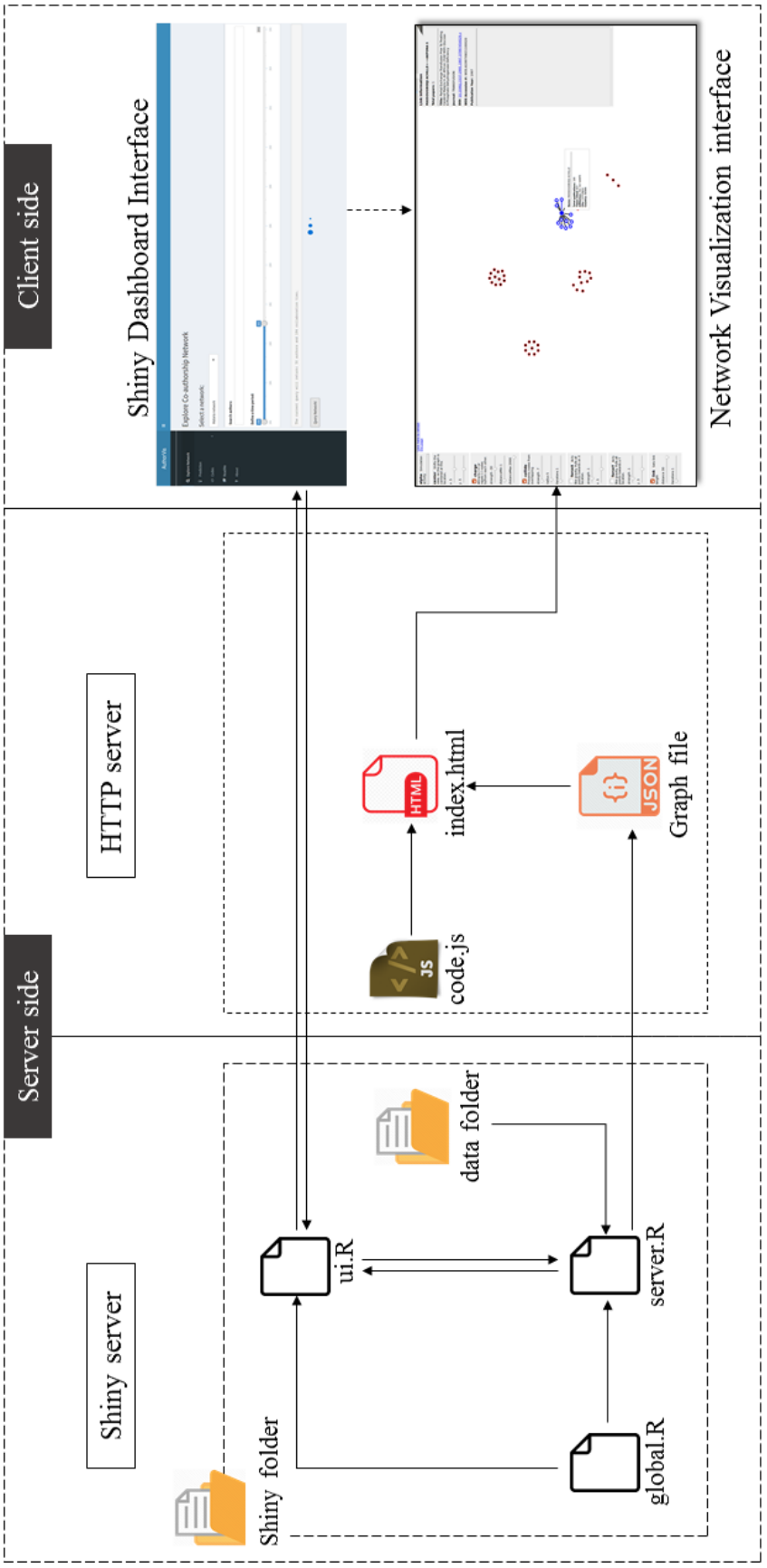


FIGURE 8.1: Authorvis Design and Architecture

Figure 8.3 is a screenshot of the dashboard prediction page with its simple query builder. It is accessible via the ”Prediction” menu option in the side bar. Here again, the user is prompted to select a network, a first and second authors, and choose a model. Upon click on the ”Predict Tie Probability!” button, the query is submitted to the server. Once the processing is done, the output is sent back to the page for display. The prediction tool is model-based and used the final ERGMs and TERGMs from chapters 5, 6, and 7 to calculate a micro-interpretation probability of collaboration between two authors [121].

8.4.2 Network Visualization Interface

The frontend Network Visualization Interface has three main parts: a left control pane, an SVG scene, and a right link information pane. The user can adjust the display of the network by modifying the default options of the physics of the network [147] using the control pane on the left. The SVG scene displays the queried network. In the SVG scene, a mouse hover over a vertex displays a tooltip of details on the author represented by the vertex, and a single click on a vertex displays a word cloud of the keywords expertise on that vertex, showing what the work of the vertex author is about. A double-click on a vertex highlights the subnetwork of the author represented by that specific vertex. Once an edge is clicked, its color turns blue and the list of published materials co-authored by the two vertices defining the clicked edge is displayed on the link information right pane. All published materials listed in the right pane can be traced back to their publication page on the web via their DOI or the WOS accession number with a single click (subfigure (d) on figure 8.2).

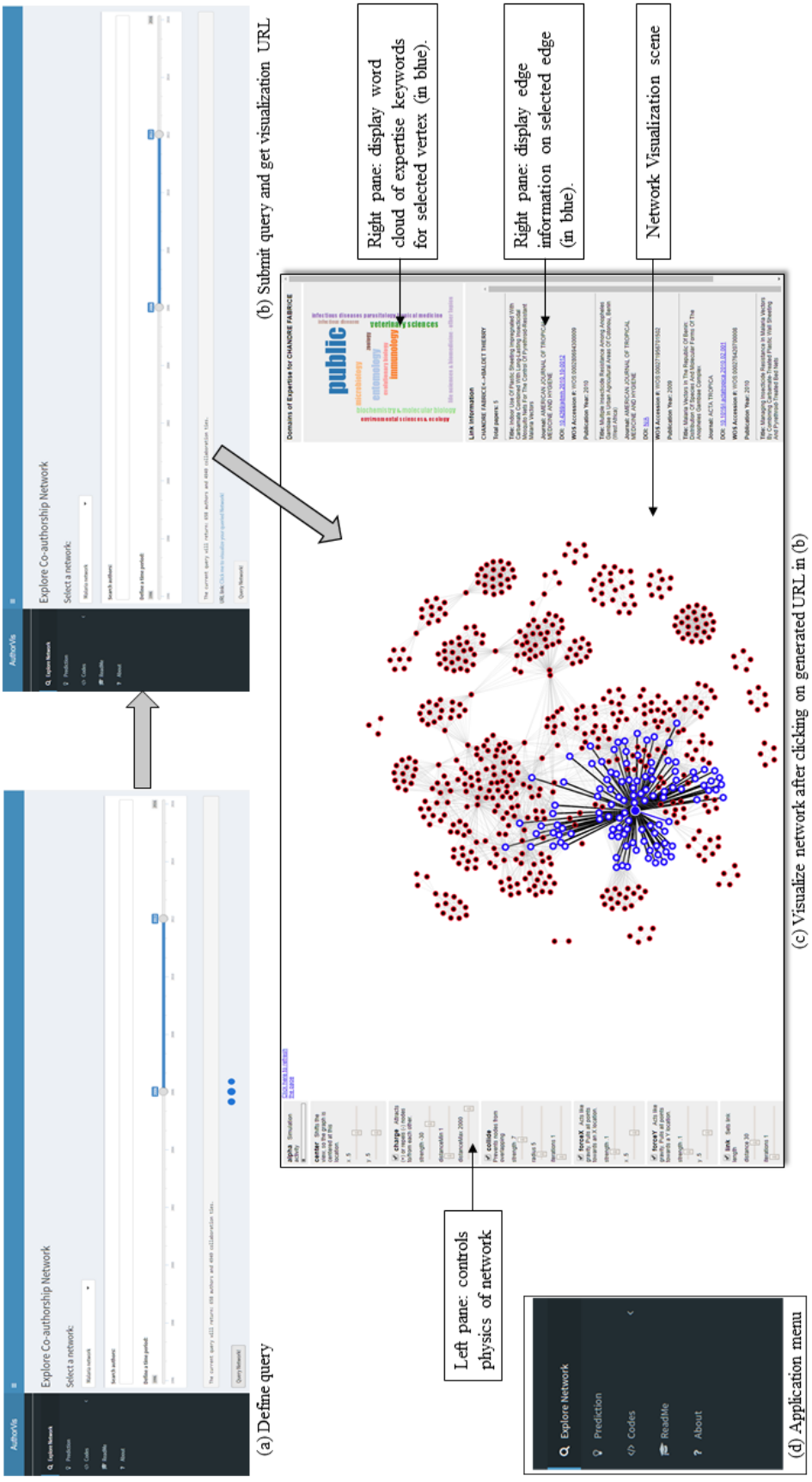


FIGURE 8.2: User View of the Authorvis co-authorship tool

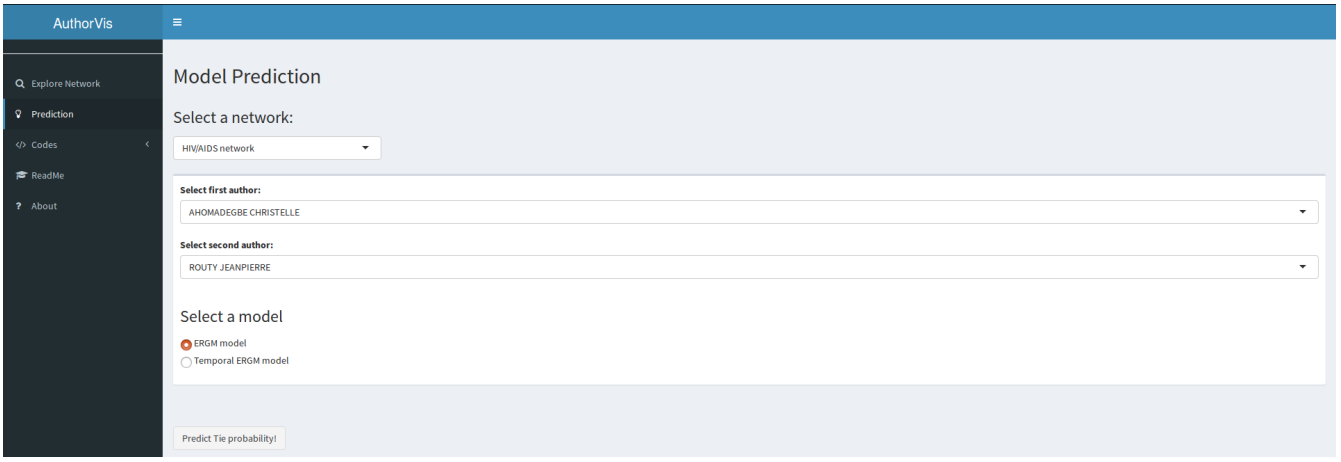


FIGURE 8.3: Screenshot of the co-authorship prediction page.

AuthorVis can be used by policy makers to visualize collaboration interactions in time between researchers. Figure 8.4, for example, depicts the co-authorship network of the 10 most cited papers in malaria research in Benin, highlighting one author (Prof. Martin AKOGBETO) as the most important author for the sustainability of the network.

8.5 Deployment

The system is packed in a Docker container to facilitate its use and installation. The docker container is accessible at `https://hub.docker.com/r/rosericazondekon/authorvis/`. The project source files can be forked or cloned from Github at `https://github.com/rosericazondekon/authorvis`.

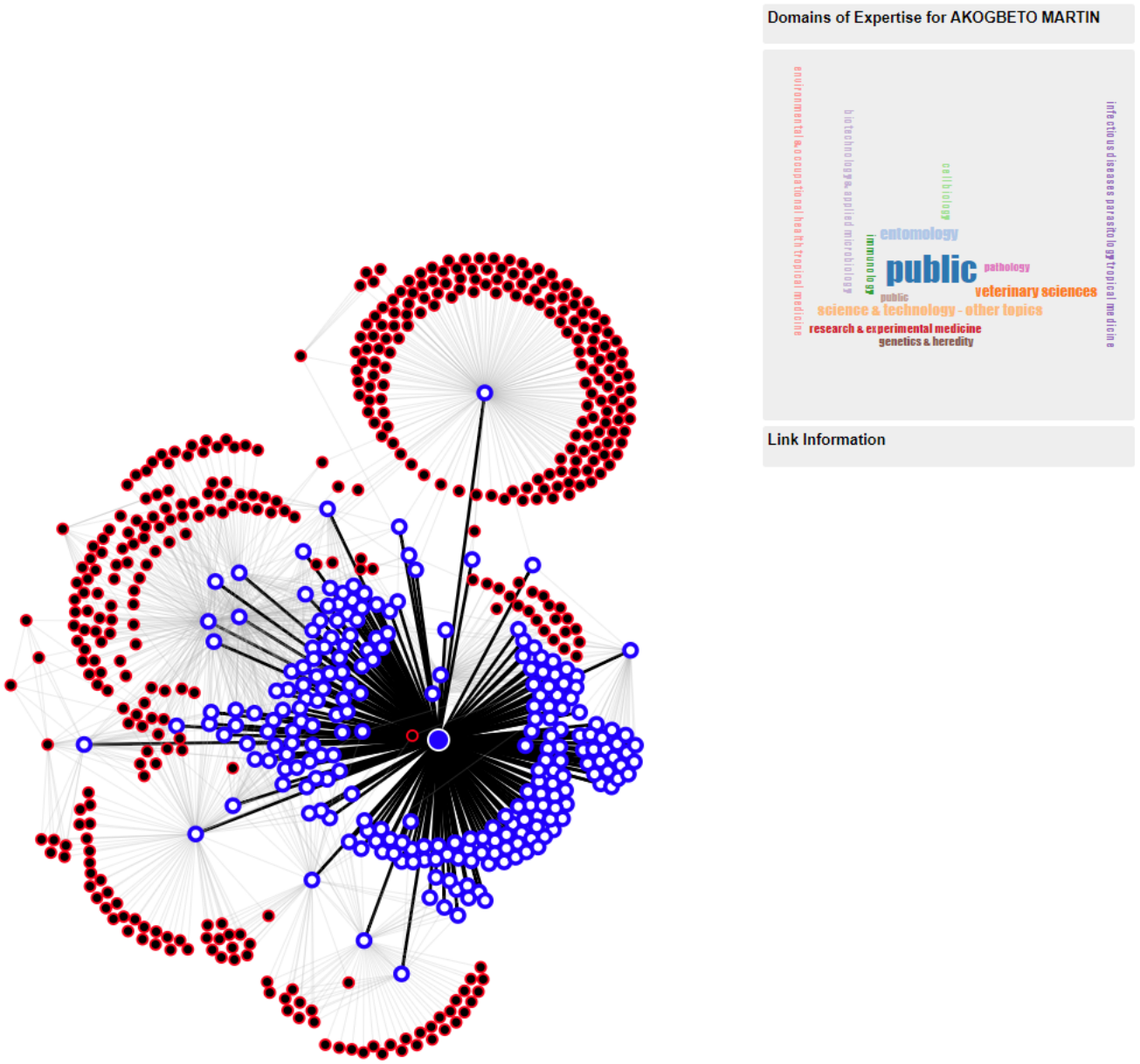


FIGURE 8.4: Co-authorship network of the top 10 most cited papers in Malaria research in Benin.

Chapter 9

General Conclusion

In this dissertation, we have documented and described the collaborative pattern in Malaria, HIV/AIDS and TB research in Benin. Our findings suggest that each one of the collaborative research network of Malaria, HIV/AIDS and TB has a complex structure. We modeled these complex structures to predict the establishment of future collaboration ties. We implemented the models in a shiny-based application for co-authorship visualization and scientific collaboration prediction tool which we named **AuthorVis**.

Strengths and Limitations

The application of temporal or dynamic modeling techniques is the major strength of our research along with its application of not only descriptive methods but also robust network analysis methods such as inferential methods like Monte-Carlo simulations, unlike most studies on co-authorship analysis. Our data mining strategy involved a robust

machine learning algorithm that helped address the crucial issue of the disambiguation of authors names and assigns a unique identifier to each of them. To the best of our knowledge, our study is the first to describe the malaria research collaborations network via co-authorship network analysis in Benin. It is also the first to apply statistical network models to investigate co-authorship networks in a specific research area in an African country.

The fact that we collected data only from the Web Of Science can be considered as an important limitation of this study. However, according to Falagas and colleagues [148], who compared PubMed, Scopus, Web Of Science and Google Scholar in their paper, the Web Of Science appears as a reasonable scientific database source for our analysis. In addition, it proved to cover a wide range of both old and recently published papers. Falagas and colleagues [148] found PubMed to be the optimal choice in terms of scientific database. For that reason we ran the same bibliographic searches in PubMed. Unfortunately, the Web Of Science returns more relevant data than PubMed.

Another major limitation is related to the manual curation of the scientific publications and the keyword based searches of the literature involved in this study. It is therefore worth acknowledging the possibility of error or incompleteness of the scientific publications reviewed. However, we limited this possibility by casting a wider net, querying the Web Of Science API with wider keywords, then narrowing the search down by combining the keywords. Yet another major limitation is that only one manual curator has reviewed the publications for the selection criteria. Having multiple curators would have allowed us to evaluate the quality of the search by measuring selection agreement statistics (kappa

statistic for example) between the curators.

The nature of all co-authorship studies in itself is another limitation of this study. Collaborators, in co-authorship networks, do not often come from the same scientific discipline, or do not play the same roles on a particular research project. The data we collected did not allow us to accurately assess or even infer the disciplines each author comes from or their specific contribution in the published documents.

Future Directions

There are several future directions. Our work can be extended to the entire African collaboration network in Malaria, HIV/AIDS and TB. Since collaborations usually are often initiated between individuals, labs or even countries, the analysis of bipartite co-authorship networks is an interesting direction to our study.

Currently, **AuthorVis** is specifically built for Malaria, TB and HIV/AIDS in Benin. Future developments may extend the tool to other research domain. Adding a general purpose module to **AuthorVis** for the visualization of any user-input co-authorship network is an interesting venture since it will also require the integration of a data preprocessing module to facilitate the disambiguation and deduplication of co-authorship information. Furthermore, incorporating a layered structured network visualization [83] functionality to the visualization in order to display temporal changes in the evolution of the co-authorship network is another interesting direction. It can, in addition be designed into a real-time, cross-domain, and cross-collection co-authorship visualization interface capable of automatically searching the literature.

Outside of the realm of co-authorship analyses, the same idea of network analyses and visualization can be extended to other important disciplines such as Neuroscience. In analogy to co-authorship networks, the brain functioning can be represented as a brain connectivity network (connectome) where parcels or anatomical regions or regions of interest of the brain represent the vertices and the edges determine statistical dependency of combined neuronal activities between the vertices.

Basic network analyses have already enabled the development of network-based clinical diagnostics of certain pathologies such as schizophrenia [149], stroke [150], and Alzheimer's disease [151]. Although trending, modeling brain connectivity networks by means of the methods used in this dissertation remains limited to very few studies [152–157] in neuroscience. Since it is important to better explain the functional organization of the brain and to allow inference of specific brain properties, the visualization of real time brain connectivity dynamics has potentials for the development of Brain Computer Interfaces. See appended neuroscience manuscript draft.

Bibliography

[1] Jonathan R Davis and Joshua Lederberg. *Emerging Infectious Diseases from the Global to the Local Perspective: Workshop Summary*. National Academies Press, 2001. ISBN 0-309-07184-4. 00021.

[2] John Luke Gallup and Jeffrey D Sachs. The economic burden of malaria. *The American journal of tropical medicine and hygiene*, 64(1 suppl):85–96, 2001. ISSN 0002-9637. 01510.

[3] M. Vitoria, R. Granich, C. F. Gilks, C. Gunneberg, M. Hosseini, W. Were, M. Raviglione, and K. M. De Cock. The Global Fight Against HIV/AIDS, Tuberculosis, and Malaria: Current Status and Future Perspectives. *American Journal of Clinical Pathology*, 131(6):844–848, June 2009. ISSN 0002-9173, 1943-7722. doi: 10.1309/AJCP5XHDB1PNAEYT. 00003.

[4] UN General Assembly. United Nations millennium declaration. *United Nations General Assembly*, 2000. 00156.

[5] Margaret Arthur. Institute for Health Metrics and Evaluation. *Nursing Standard*, 28(42):32–32, 2014. ISSN 0029-6570. 00000.

[6] Christopher J L Murray, Katrina F Ortblad, Caterina Guinovart, Stephen S Lim, Timothy M Wolock, D Allen Roberts, Emily A Dansereau, Nicholas Graetz, Ryan M Barber, Jonathan C Brown, Haidong Wang, Herbert C Duber, Mohsen Naghavi, Daniel Dicker, Lalit Dandona, Joshua A Salomon, Kyle R Heuton, Kyle Foreman, David E Phillips, Thomas D Fleming, Abraham D Flaxman, Bryan K Phillips, Elizabeth K Johnson, Megan S Coggeshall, Foad Abd-Allah, Semaw Ferede Abera, Jerry P Abraham, Ibrahim Abubakar, Laith J Abu-Raddad, Niveen Me Abu-Rmeileh, Tom Achoki, Austine Olufemi Adeyemo, Arsène Kouablan Adou, José C Adsuar, Emilie Elisabet Agardh, Dickens Akena, Mazin J Al Kahbouri, Deena Alasfoor, Mohammed I Albittar, Gabriel Alcalá-Cerra, Miguel Angel Alegretti, Zewdie Aderaw Alemu, Rafael Alfonso-Cristancho, Samia Alhabib, Raghib Ali, Francois Alla, Peter J Allen, Ubai Alsharif, Elena Alvarez, Nelson Alvis-Guzman, Adansi A Amankwaa, Azmeraw T Amare, Hassan Amini, Walid Ammar, Benjamin O Anderson, Carl Abelardo T Antonio, Palwasha Anwari, Johan Ärnlöv, Valentina S Arsic Arsenijevic, Ali Artaman, Rana J Asghar, Reza Assadi, Lydia S Atkins, Alaa Badawi, Kalpana Balakrishnan, Amitava Banerjee, Sanjay Basu, Justin Beardsley, Tolesa Bekele, Michelle L Bell, Eduardo Bernabe, Tariku Jibat Beyene, Neeraj Bhala, Ashish Bhalla, Zulfiqar A Bhutta, Aref Bin Abdulhak, Agnes Binagwaho, Jed D Blore, Berrak Bora Basara, Dipan Bose, Michael Brainin, Nicholas Breitborde, Carlos A Castañeda-Orjuela, Ferrán Catalá-López, Vineet K Chadha, Jung-Chen Chang, Peggy Pei-Chia Chiang, Ting-Wu Chuang, Mercedes Colomar, Leslie Trumbull

Cooper, Cyrus Cooper, Karen J Courville, Benjamin C Cowie, Michael H Criqui, Rakhi Dandona, Anand Dayama, Diego De Leo, Louisa Degenhardt, Borja Del Pozo-Cruz, Kebede Deribe, Don C Des Jarlais, Muluken Dessalegn, Samath D Dharmaratne, Uğur Dilmen, Eric L Ding, Tim R Driscoll, Adnan M Durrani, Richard G Ellenbogen, Sergey Petrovich Ermakov, Alireza Esteghamati, Emerito Jose A Faraon, Farshad Farzadfar, Seyed-Mohammad Fereshtehnejad, Daniel Obadare Fijabi, Mohammad H Forouzanfar, Urbano Fra.Paleo, Lynne Gaffikin, Amiran Gamkrelidze, Fortuné Gbètoho Gankpé, Johanna M Geleijnse, Bradford D Gessner, Katherine B Gibney, Ibrahim Abdelmageem Mohamed Ginawi, Elizabeth L Glaser, Philimon Gona, Atsushi Goto, Hebe N Gouda, Harish Chander Gugnani, Rajeev Gupta, Rahul Gupta, Nima Hafezi-Nejad, Randah Ribhi Hamadeh, Mouhanad Hammami, Graeme J Hankey, Hilda L Harb, Josep Maria Haro, Rasmus Havmoeller, Simon I Hay, Mohammad T Hedayati, Ileana B Heredia Pi, Hans W Hoek, John C Hornberger, H Dean Hosgood, Peter J Hotez, Damian G Hoy, John J Huang, Kim M Iburg, Bulat T Idrisov, Kaire Innos, Kathryn H Jacobsen, Panniyammakal Jeemon, Paul N Jensen, Vivekanand Jha, Guohong Jiang, Jost B Jonas, Knud Juel, Haidong Kan, Ida Kankindi, Nadim E Karam, André Karch, Corine Kakizi Karema, Anil Kaul, Norito Kawakami, Dhruv S Kazi, Andrew H Kemp, Andre Pascal Kengne, Andre Keren, Maia Kereselidze, Yousef Saleh Khader, Shams Eldin Ali Hassan Khalifa, Ejaz Ahmed Khan, Young-Ho Khang, Irma Khonelidze, Yohannes Kinfu, Jonas M Kinge, Luke Knibbs, Yoshihiro Kokubo, S Kosen, Barthelemy Kuate Defo, Veena S Kulkarni,

Chanda Kulkarni, Kaushalendra Kumar, Ravi B Kumar, G Anil Kumar, Gene F Kwan, Taavi Lai, Arjun Lakshmana Balaji, Hilton Lam, Qing Lan, Van C Lansingh, Heidi J Larson, Anders Larsson, Jong-Tae Lee, James Leigh, Mall Leinsalu, Ricky Leung, Yichong Li, Yongmei Li, Graça Maria Ferreira De Lima, Hsien-Ho Lin, Steven E Lipshultz, Shiwei Liu, Yang Liu, Belinda K Lloyd, Paulo A Lotufo, Vasco Manuel Pedro Machado, Jennifer H Maclachlan, Carlos Magis-Rodriguez, Marek Majdan, Christopher Chabila Mapoma, Wagner Marcenes, Melvin Barrientos Marzan, Joseph R Masci, Mohammad Taufiq Mashal, Amanda J Mason-Jones, Bongani M Mayosi, Tasara T Mazorodze, Abigail Cecilia Mckay, Peter A Meaney, Man Mohan Mehndiratta, Fabiola Mejia-Rodriguez, Yohannes Adama Melaku, Ziad A Memish, Walter Mendoza, Ted R Miller, Edward J Mills, Karzan Abdulmuhsin Mohammad, Ali H Mokdad, Glen Liddell Mola, Lorenzo Monasta, Marcella Montico, Ami R Moore, Rintaro Mori, Wilkister Nyaora Moturi, Mitsuru Mukaigawara, Kinnari S Murthy, Aliya Naheed, Kovin S Naidoo, Luigi Naldi, Vinay Nangia, K M Venkat Narayan, Denis Nash, Chakib Nejjari, Robert G Nelson, Sudan Prasad Neupane, Charles R Newton, Marie Ng, Muhammad Imran Nisar, Sandra Nolte, Ole F Norheim, Vincent Nowaseb, Luke Nyakarahuka, In-Hwan Oh, Takayoshi Ohkubo, Bolajoko O Olusanya, Saad B Omer, John Nelson Opio, Orish Ebere Orisakwe, Jeyaraj D Pandian, Christina Papachristou, Angel J Paternina Caicedo, Scott B Patten, Vinod K Paul, Boris Igor Pavlin, Neil Pearce, David M Pereira, Aslam Pervaiz, Konrad Pesudovs, Max Petzold, Farshad Pourmalek, Dima Qato,

Amado D Quezada, D Alex Quistberg, Anwar Rafay, Kazem Rahimi, Vafa Rahimi-Movaghar, Sajjad Ur Rahman, Murugesan Raju, Saleem M Rana, Homie. Global, regional, and national incidence and mortality for HIV, tuberculosis, and malaria during 1990–2013: A systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*, 384(9947):1005–1070, September 2014. ISSN 01406736. doi: 10.1016/S0140-6736(14)60844-8. 00405.

[7] Craig Stoops. President's Malaria Initiative. Technical report, DTIC Document, 2008. 00000.

[8] Global Fund. Making a difference: Global fund results report 2011. *The Global Fund, Geneva*, 2011. 00005.

[9] World Health Organization. World malaria report 2010. *Geneva: World Health Organization View Article Google Scholar*, 2012. 00161.

[10] Lawrence M Barat. Four malaria success stories: How malaria burden was successfully reduced in Brazil, Eritrea, India, and Vietnam. *The American journal of tropical medicine and hygiene*, 74(1):12–16, 2006. ISSN 0002-9637. 00117.

[11] Martin C. Akogbéto, Rock Y. Aïkpon, Roseric Azondékon, Gil G. Padonou, Razaki A. Ossè, Fiacre R. Agossa, Raymond Beach, and Michel Sèzonlin. Six years of experience in entomological surveillance of indoor residual spraying against malaria transmission in Benin: Lessons learned, challenges and outlooks. *Malaria Journal*, 14(1), December 2015. ISSN 1475-2875. doi: 10.1186/s12936-015-0757-5. 00002.

[12] Joint United Nations Programme on HIV/AIDS. Getting to Zero: 2011–2015 strategy. 2010. 00035.

[13] Joint United Nations Programme on HIV/AIDS. *Global Report: UNAIDS Report on the Global AIDS Epidemic 2010*. UNAIDS, 2010. ISBN 92-9173-871-9. 01036.

[14] World Health Organization. *Global Tuberculosis Control: WHO Report 2010*. World Health Organization, 2010. ISBN 92-4-156406-7. 00010.

[15] World Health Organization. Economic costs of malaria are many times higher than previously estimated. In *Economic Costs of Malaria Are Many Times Higher than Previously Estimated*. 2000. 00012.

[16] Linda M. Richter, Knut Lönnroth, Chris Desmond, Robin Jackson, Ernesto Jaramillo, and Diana Weil. Economic Support to Patients in HIV and TB Grants in Rounds 7 and 10 from the Global Fund to Fight AIDS, Tuberculosis and Malaria. *PLoS ONE*, 9(1):e86225, January 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0086225. 00015.

[17] Dean T Jamison. *Disease and Mortality in Sub-Saharan Africa*. World Bank Publications, 2006. ISBN 0-8213-6398-0. 00259.

[18] Carole A Long and Fidel Zavala. Malaria vaccines and human immune responses. *Current Opinion in Microbiology*, 32:96–102, August 2016. ISSN 13695274. doi: 10.1016/j.mib.2016.04.006. 00003.

[19] Fausto Titti, Aurelio Cafaro, Flavia Ferrantelli, Antonella Tripiciano, Sonia Moretti, Antonella Caputo, Riccardo Gavioli, Fabrizio Ensoli, Marjorie Robert-Guroff, Susan Barnett, and Barbara Ensoli. Problems and emerging approaches in HIV/AIDS vaccine development. *Expert Opinion on Emerging Drugs*, 12(1):23–48, March 2007. ISSN 1472-8214, 1744-7623. doi: 10.1517/14728214.12.1.23. 00036.

[20] B. D. Walker and D. R. Burton. Toward an AIDS Vaccine. *Science*, 320(5877): 760–764, May 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1152622. 00407.

[21] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, January 2001. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.98.2.404. 04061.

[22] Elizabeth L. Corbett, Catherine J. Watt, Neff Walker, Dermot Maher, Brian G. Williams, Mario C. Raviglione, and Christopher Dye. The Growing Burden of Tuberculosis: Global Trends and Interactions With the HIV Epidemic. *Archives of Internal Medicine*, 163(9):1009, May 2003. ISSN 0003-9926. doi: 10.1001/archinte.163.9.1009. 02947.

[23] Neel R. Gandhi, N. Sarita Shah, Jason R. Andrews, Venanzio Vella, Anthony P. Moll, Michelle Scott, Darren Weissman, Claudio Marra, Umesh G. Lalloo, and Gerald H. Friedland. HIV Coinfection in Multidrug- and Extensively Drug-Resistant Tuberculosis Results in High Early Mortality. *American Journal*

of Respiratory and Critical Care Medicine, 181(1):80–86, January 2010. ISSN 1073-449X, 1535-4970. doi: 10.1164/rccm.200907-0989OC. 00248.

[24] H.B. Ghafouri, H. Mohammadhassanzadeh, F. Shokraneh, M. Vakilian, and S. Farahmand. Social network analysis of Iranian researchers on emergency medicine: A sociogram analysis. *Emergency Medicine Journal*, 31(8):619–624, 2014. doi: 10.1136/emermed-2012-201781. 00008.

[25] Carlos Medicis Morel, Suzanne Jacob Serruya, Gerson Oliveira Penna, and Reinaldo Guimarães. Co-authorship Network Analysis: A Powerful Tool for Strategic Planning of Research, Development and Capacity Building Programs on Neglected Diseases. *PLoS Neglected Tropical Diseases*, 3(8):e501, August 2009. ISSN 1935-2735. doi: 10.1371/journal.pntd.0000501. 00096.

[26] Gregorio González-Alcaide, Jinseo Park, Charles Huamaní, Joaquín Gascón, and José Manuel Ramos. Scientific authorships and collaboration network analysis on Chagas disease: Papers indexed in PubMed (1940-2009). *Revista do Instituto de Medicina Tropical de São Paulo*, 54(4):219–228, 2012. ISSN 0036-4665. 00028.

[27] Sam M Mbulaiteye, Kishor Bhatia, Clement Adebamowo, and Annie J Sasco. HIV and cancer in Africa: Mutual collaboration between HIV and cancer programs may provide timely research and public health data. *Infectious Agents and Cancer*, 6(1):16, 2011. ISSN 1750-9378. doi: 10.1186/1750-9378-6-16. 00038.

[28] U. D'Alessandro, B.O. Olaleye, W. McGuire, M.C. Thomson, P. Langerock, S. Bennett, and B.M. Greenwood. A comparison of the efficacy of

insecticide-treated and untreated bed nets in preventing malaria in Gambian children. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 89 (6):596–598, November 1995. ISSN 00359203. doi: 10.1016/0035-9203(95)90401-8.

[29] Joint United Nations Programme on HIV/AIDS and World Health Organization. *AIDS Epidemic Update, December 2006*. World Health Organization, 2007. ISBN 92-9173-542-6. 00519.

[30] Alan Whiteside. *HIV/AIDS: A Very Short Introduction*, volume 174. Oxford University Press, 2008. ISBN 0-19-280692-0. 00119.

[31] Centers for Disease Control and Prevention. Revised recommendations for HIV testing of adults, adolescents, and pregnant women in health-care settings. *Annals of Emergency Medicine*, 49(5):575–577, 2007. ISSN 0196-0644. 08464.

[32] Bluma Brenner and Mark A. Wainberg. We need to use the best antiretroviral drugs worldwide to prevent HIV drug resistance:. *AIDS*, 30(17):2725–2727, November 2016. ISSN 0269-9370. doi: 10.1097/QAD.0000000000001234. 00000.

[33] Alexandra Calmy, Fernando Pascual, and Nathan Ford. HIV drug resistance. *New England Journal of Medicine*, 350(26):2720–2721, 2004. ISSN 0028-4793. 00038.

[34] François Clavel and Allan J Hance. HIV drug resistance. *New England Journal of Medicine*, 350(10):1023–1035, 2004. ISSN 0028-4793. 00817.

[35] Stefan HE Kaufmann and Paul van Helden. *Handbook of Tuberculosis: Clinics, Diagnostics, Therapy and Epidemiology*. Wiley-VCH, 2008. ISBN 3-527-31888-7. 00010.

[36] Alimuddin Zumla. Handbook of tuberculosis. *The Lancet Infectious Diseases*, 9 (12):736, 2009. ISSN 1473-3099. 00000.

[37] MC Raviglione, AD Harries, R Msiska, David Wilkinson, and P Nunn. Tuberculosis and HIV: Current status in Africa. *AIDS (London, England)*, 11: S115, 1997. ISSN 0269-9370. 00252.

[38] SK Sharma, Alladi Mohan, and Tamilarasu Kadhiravan. HIV-TB co-infection: Epidemiology, diagnosis & management. *Indian Journal of Medical Research*, 121 (4):550–567, 2005. ISSN 0971-5916.

[39] Z Toossi, Hirsch Mayanja-Kizza, CS Hirsch, KL Edmonds, T Spahlinger, DL Hom, H Aung, P Mugyenyi, JJ Ellner, and CW Whalen. Impact of tuberculosis (TB) on HIV-1 activity in dually infected patients. *Clinical & Experimental Immunology*, 123(2):233–238, 2001. ISSN 1365-2249. 00175.

[40] Lia D'Anibrosio, Antonio Spanevello, and Rosella Centis. Epidemiology of TB. *Tuberculosis*, 58:14, 2014. ISSN 1849840288. 00000.

[41] Centers for Disease Control and Prevention (CDC). Emergence of Mycobacterium tuberculosis with extensive resistance to second-line drugs–worldwide, 2000-2004. *MMWR. Morbidity and mortality weekly report*, 55(11):301, 2006. ISSN 1545-861X. 00651.

[42] World Health Organization. Multidrug and extensively drug-resistant TB. 2010. 00915.

[43] Robert L Cowie. The epidemiology of tuberculosis in gold miners with silicosis. *American journal of respiratory and critical care medicine*, 150(5):1460–1462, 1994. ISSN 1073-449X. 00176.

[44] Emmanuel M Mulenga, Hugh B Miller, Thomson Sinkala, Tracy A Hysong, and Jefferey L Burgess. Silicosis and tuberculosis in Zambian miners. *International journal of occupational and environmental health*, 2013. 00014.

[45] D Rees and J Murray. Silica, silicosis and tuberculosis [State of the Art Series. Occupational lung disease in high-and low-income countries, Edited by M. Chan-Yeung. Number 4 in the series]. *The International Journal of Tuberculosis and Lung Disease*, 11(5):474–484, 2007. ISSN 1027-3719. 00132.

[46] Marianne E Sinka, Michael J Bangs, Sylvie Manguin, Yasmin Rubio-Palis, Theeraphap Chareonviriyaphap, Maureen Coetzee, Charles M Mbogo, Janet Hemingway, Anand P Patil, and William H Temperley. A global map of dominant malaria vectors. *Parasites & vectors*, 5(1):1, 2012. ISSN 1756-3305. 00189.

[47] Robert W. Snow, Carlos A. Guerra, Abdisalan M. Noor, Hla Y. Myint, and Simon I. Hay. The global distribution of clinical episodes of Plasmodium falciparum malaria. *Nature*, 434(7030):214–217, March 2005. ISSN 0028-0836, 1476-4679. doi: 10.1038/nature03342. 02814.

[48] S. P. James and P. Tate. New Knowledge of the Life-Cycle of Malaria Parasites. *Nature*, 139(3517):545–545, March 1937. ISSN 0028-0836. doi: 10.1038/139545a0. 00080.

[49] P.L. Alonso, S.W. Lindsay, J.R.M.Armstrong Schellenberg, K. Keita, P. Gomez, F.C. Shenton, A.G. Hill, P.H. David, G. Fegan, K. Cham, and B.M. Greenwood. A malaria control trial using insecticide-treated bed nets and targeted chemoprophylaxis in a rural area of The Gambia, West Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 87:37–44, June 1993. ISSN 00359203. doi: 10.1016/0035-9203(93)90174-O.

[50] Katherine E. Battle, Donal Bisanzio, Harry S. Gibson, Samir Bhatt, Ewan Cameron, Daniel J. Weiss, Bonnie Mappin, Ursula Dalrymple, Rosalind E. Howes, Simon I. Hay, and Peter W. Gething. Treatment-seeking rates in malaria endemic countries. *Malaria Journal*, 15(1), December 2016. ISSN 1475-2875. doi: 10.1186/s12936-015-1048-x.

[51] Loet Leydesdorff and Staša Milojević. Scientometrics. *arXiv preprint arXiv:1208.4566*, 2012. 00467.

[52] Garfield Eugene. Citation Indexing, Its Theory and Application in Science, Technology, and Humanities. 1979.

[53] Terttu Luukkonen, Olle Persson, and Gunnar Sivertsen. Understanding patterns of international scientific collaboration. *Science, Technology & Human Values*, 17 (1):101–126, 1992. ISSN 0162-2439. 00457.

[54] Caroline S. Wagner. Six case studies of international collaboration in science. *Scientometrics*, 62(1):3–26, January 2005. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-005-0001-0. 00193.

[55] Wolfgang Glänzel and András Schubert. Analysing scientific networks through co-authorship. In *Handbook of Quantitative Science and Technology Research*, pages 257–276. Springer, 2004. 00493.

[56] M. E. J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5200–5205, April 2004. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0307545100. 01352.

[57] M. E. J. Newman. Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64(1), June 2001. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.64.016131.

[58] M. E. J. Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1), June 2001. ISSN 1063-651X, 1095-3787. doi: 10.1103/PhysRevE.64.016132. 02213.

[59] Katy Börner, Chaomei Chen, and Kevin W. Boyack. Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255, January 2003. ISSN 1550-8382. doi: 10.1002/aris.1440370106. 00981.

[60] Andrea Scharnhorst, Katy Börner, and Peter van den Besselaar, editors. *Models of Science Dynamics: Encounters between Complexity Theory and Information*

Sciences. Understanding complex systems. Springer, Heidelberg ; New York, 2012. ISBN 978-3-642-23067-7. 00050.

[61] F. Mali, L. Kronegger, P. Doreian, and A. Ferligoj. *Dynamic Scientific Co-Authorship Networks*. Understanding Complex Systems. 2012. 00050.

[62] Helga Bermeo Andrade, Ernesto de los Reyes López, and Tomas Bonavia Martín. Dimensions of scientific collaboration and its contribution to the academic research groups' scientific quality. *Research Evaluation*, 18(4):301–311, October 2009. ISSN 09582029, 14715449. doi: 10.3152/095820209X451041. 00038.

[63] Juan D Rogers, Barry Bozeman, and Ivan Chompalov. Obstacles and opportunities in the application of network analysis to the evaluation of R&D. *Research Evaluation*, 10(3):161–172, December 2001. ISSN 09582029, 14715449. doi: 10.3152/147154401781777033. 00000.

[64] Diane H. Sonnenwald. Scientific collaboration. *Annual Review of Information Science and Technology*, 41(1):643–681, 2007. ISSN 00664200. doi: 10.1002/aris.2007.1440410121. 00418.

[65] Haiyan Hou, Hildrun Kretschmer, and Zeyuan Liu. The structure of scientific collaboration networks in Scientometrics. *Scientometrics*, 75(2):189–202, May 2008. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-007-1771-3. 00186.

[66] Gregorio González-Alcaide, Rafael Aleixandre-Benavent, Carolina Navarro-Molina, and Juan Carlos Valderrama-Zurián. Coauthorship networks and

institutional collaboration patterns in reproductive biology. *Fertility and Sterility*, 90(4):941–956, October 2008. ISSN 00150282. doi: 10.1016/j.fertnstert.2007.07.1378. 00035.

[67] Hannes Toivanen and Branco Ponomariov. African regional innovation systems: Bibliometric analysis of research collaboration patterns 2005–2009. *Scientometrics*, 88(2):471–493, August 2011. ISSN 0138-9130, 1588-2861. doi: 10.1007/s11192-011-0390-1. 00035.

[68] L. Bellanca. Measuring interdisciplinary research: Analysis of co-authorship for research staff at the University of York. *Bioscience Horizons*, 2(2):99–112, June 2009. ISSN 1754-7431. doi: 10.1093/biohorizons/hzp012.

[69] R. Aleixandre-Benavent, G. González-Alcaide, A. Alonso-Arroyo, M. Bolaños-Pizarro, L. Castelló-Cogollos, and J.C. Valderrama-Zurián. Coauthorship Networks and Institutional Collaboration in Farmacia Hospitalaria. *Farmacia Hospitalaria (English Edition)*, 32(4):226–233, January 2008. ISSN 21735085. doi: 10.1016/S2173-5085(08)70044-3. 00003.

[70] P. Salamati and F. Soheili. Social network analysis of Iranian researchers in the field of violence. *Chinese Journal of Traumatology - English Edition*, 19(5): 264–270, 2016. doi: 10.1016/j.cjtee.2016.06.008. 00000.

[71] F. Sadoughi, A. Valinejadi, M. Serati Shirazi, and R. Khademi. Social network analysis of Iranian researchers on medical parasitology: A 41 year co-authorship survey. *Iranian Journal of Parasitology*, 11(2):204–212, 2016. 00001.

[72] Vladimir Batagelj and Andrej Mrvar. Pajek. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*, pages 1245–1256. Springer New York, New York, NY, 2014. ISBN 978-1-4614-6169-2 978-1-4614-6170-8. doi: 10.1007/978-1-4614-6170-8_310.

[73] Stephen P. Borgatti, Martin G. Everett, and Linton C. Freeman. UCINET. In Reda Alhajj and Jon Rokne, editors, *Encyclopedia of Social Network Analysis and Mining*, pages 2261–2267. Springer New York, New York, NY, 2014. ISBN 978-1-4614-6169-2 978-1-4614-6170-8. doi: 10.1007/978-1-4614-6170-8_316.

[74] Qing Zhang. *Complex Network Analysis for Scientific Collaboration Prediction and Biological Hypothesis Generation*. PhD thesis, University of Wisconsin-Milwaukee, Milwaukee, WI, USA, 2014. 00000.

[75] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009. ISSN 1931-0145.

[76] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[77] Sandra Cristina Oliveira, Juliana Cobre, and Taiane de Paula Ferreira. A Bayesian approach for the reliability of scientific co-authorship networks with

emphasis on nodes. *Social Networks*, 48:110–115, January 2017. ISSN 0378-8733. doi: 10.1016/j.socnet.2016.06.005. 00000.

[78] Xiaoming Liu, Johan Bollen, Michael L. Nelson, Herbert Van de Sompel, Jeremy Hussell, Rick Luce, and Linn Marks. Toolkits for visualizing co-authorship graph. page 404. ACM Press, 2004. ISBN 978-1-58113-832-0. doi: 10.1145/996350.996470.

[79] Eduardo M. Barbosa, Mirella M. Moro, Giseli Rabello Lopes, and J. Palazzo M. de Oliveira. VRRC: Web based tool for visualization and recommendation on co-authorship network (abstract only). page 865. ACM Press, 2012. ISBN 978-1-4503-1247-9. doi: 10.1145/2213836.2213975.

[80] Fabian Odoni, Wolfgang Semar, and Elena Mastrandrea. Visualisation of Collaboration in Social Collaborative Knowledge Management Systems. In *Understanding Information Spaces. Proceedings of the 15th International Symposium of Information Science (ISI 2017)*, pages 386–388, Berlin, 13-15 March 2017.

[81] Miguel Grinberg. *Flask Web Development: Developing Web Applications with Python*. " O'Reilly Media, Inc.", 2014. ISBN 1-4919-4761-6.

[82] Michael Bostock. D3. js. *Data Driven Documents*, 492:701, 2012.

[83] Nagayoshi Nakazono, Kazuo Misue, and Jiro Tanaka. NeL 2: Network drawing tool for handling layered structured network diagram. pages 109–115. Australian Computer Society, Inc., 2006. ISBN 1-920682-41-4.

[84] Masashi Toyoda and Masaru Kitsuregawa. A system for visualizing and analyzing the evolution of the web with a time series of graphs. pages 151–160. ACM, 2005. ISBN 1-59593-168-6.

[85] Chaomei Chen and Leslie Carr. Visualizing the evolution of a subject domain: A case study. pages 449–452. IEEE Computer Society Press, 1999. ISBN 0-7803-5897-X.

[86] Cesim Erten, Stephen G Kobourov, Vu Le, and Armand Navabi. Simultaneous Graph Drawing: Layout Algorithms and Visualization Schemes. *J. Graph ALgorithms Appl.*, 9(1):165–182, 2005.

[87] Hui Han, Hongyuan Zha, and C. Lee Giles. Name disambiguation in author citations using a K-way spectral clustering method. page 334. ACM Press, 2005. ISBN 978-1-58113-876-4. doi: 10.1145/1065385.1065462.

[88] Vetle I. Torvik and Neil R. Smalheiser. Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data*, 3(3):1–29, July 2009. ISSN 15564681. doi: 10.1145/1552303.1552304.

[89] Pedro DeRose, Warren Shen, Fei Chen, Yoonkyong Lee, Douglas Burdick, AnHai Doan, and Raghu Ramakrishnan. DBLife: A community information management platform for the database research community. pages 169–172, 2007.

[90] Laurel L. Haak, Martin Fenner, Laura Paglione, Ed Pentz, and Howard Ratner. ORCID: A system to uniquely identify researchers. *Learned Publishing*, 25(4): 259–264, October 2012. ISSN 09531513, 17414857. doi: 10.1087/20120404.

[91] Neil R. Smalheiser and Vetle I. Torvik. Author name disambiguation. *Annual Review of Information Science and Technology*, 43(1):1–43, 2009. ISSN 00664200. doi: 10.1002/aris.2009.1440430113.

[92] Anderson A Ferreira, Marcos André Gonçalves, and Alberto HF Laender. A brief survey of automatic methods for author name disambiguation. *Acm Sigmod Record*, 41(2):15–26, 2012. ISSN 0163-5808. 00124.

[93] C Lee Giles, Hongyuan Zha, and Hui Han. Name disambiguation in author citations using a k-way spectral clustering method. pages 334–343. IEEE, 2005. ISBN 1-58113-876-8. 00277.

[94] Mikhail Yuryevich Bilenko. *Learnable Similarity Functions and Their Application to Record Linkage and Clustering*. PhD thesis, University of Texas at Austin, Austin, TX, USA, 2006. 00019.

[95] Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977. ISSN 0038-0431. 05656.

[96] Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972. ISSN 0022-250X. 01823.

[97] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. ISSN 0033-3123. 01961.

[98] Eric D Kolaczyk and Gábor Csárdi. *Statistical Analysis of Network Data with R*, volume 65. Springer, 2014. 00792.

[99] Paul Erdös and Alfréd Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959. 03506.

[100] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960. 03290.

[101] Paul Erdős and Alfréd Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Academiae Scientiarum Hungarica*, 12(1-2):261–267, 1964. ISSN 0001-5954. 00797.

[102] Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4): 1141–1144, 1959. ISSN 0003-4851. 00932.

[103] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998. ISSN 0028-0836. 32864.

[104] Vera Van Noort, Berend Snel, and Martijn A Huynen. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, 5(3):280–284, 2004. ISSN 1469-221X. 00213.

[105] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. ISSN 0036-8075. 27102.

[106] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Internet: Diameter of the world-wide web. *nature*, 401(6749):130–131, 1999. ISSN 0028-0836. 05099.

[107] Hawoong Jeong, Zoltan Néda, and Albert-László Barabási. Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)*, 61(4):567, 2003. ISSN 0295-5075. 00487.

[108] François Lorrain and Harrison C. White. Structural equivalence of individuals in social networks. *The Journal of Mathematical Sociology*, 1(1):49–80, January 1971. ISSN 0022-250X, 1545-5874. doi: 10.1080/0022250X.1971.9989788.

[109] Patrick Doreian, Vladimir Batagelj, and Anuska Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, Cambridge, 2004. ISBN 978-0-511-58417-6. doi: 10.1017/CBO9780511584176.

[110] Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social networks*, 29(2): 173–191, 2007. ISSN 0378-8733. 00000.

[111] Steve Hanneke, Wenjie Fu, and Eric P Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010. ISSN 1935-7524.

[112] Garry Robins and Philippa Pattison. Random graph models for temporal processes in social networks. *Journal of Mathematical Sociology*, 25(1):5–41, 2001. ISSN 0022-250X.

[113] Philip Leifeld, Skyler J Cranmer, and Bruce A Desmarais. Temporal Exponential Random Graph Models with xergm: Estimation and Bootstrap Confidence Intervals. *Journal of Statistical Software*, 2015. 00000.

[114] Daniel A Schult and P Swart. Exploring network structure, dynamics, and function using NetworkX. volume 2008, pages 11–16, 2008. 01142.

[115] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008. ISSN 0960-3174, 1573-1375. doi: 10.1007/s11222-007-9046-7.

[116] Hugo Zanghi, Christophe Ambroise, and Vincent Miele. Fast online graph clustering via Erdős–Rényi mixture. *Pattern Recognition*, 41(12):3592–3599, December 2008. ISSN 00313203. doi: 10.1016/j.patcog.2008.06.019.

[117] Hugo Zanghi, Franck Picard, Vincent Miele, and Christophe Ambroise. Strategies for online inference of model-based clustering in large and growing networks. *The Annals of Applied Statistics*, 4(2):687–714, June 2010. ISSN 1932-6157. doi: 10.1214/10-AOAS359.

[118] P Latouche, E Birmelé, and C Ambroise. Variational Bayesian inference and complexity control for stochastic block models. *Statistical Modelling: An International Journal*, 12(1):93–115, February 2012. ISSN 1471-082X, 1477-0342. doi: 10.1177/1471082X1001200105.

[119] MS Handcock, DR Hunter, CT Butts, SM Goodreau, and M Morris. ERGM: Fit, simulate and diagnose exponential-family models for networks, Version 2.1. *URL: http://statnetproject. org*, 2003.

[120] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. Ergm: A Package to Fit, Simulate and Diagnose

Exponential-Family Models for Networks. *Journal of Statistical Software*, 24(3): nihpa54860, May 2008. ISSN 1548-7660.

[121] Bruce A. Desmarais and Skyler J. Cranmer. Micro-Level Interpretation of Exponential Random Graph Models with Application to Estuary Networks: Desmarais/Cranmer: Micro-Level Interpretation of ERGM. *Policy Studies Journal*, 40(3):402–434, August 2012. ISSN 0190292X. doi: 10.1111/j.1541-0072.2012.00459.x.

[122] Peter Hoff. Modeling homophily and stochastic equivalence in symmetric relational data. pages 657–664, 2008.

[123] Peter Hoff. Eigenmodel: Semiparametric factor and regression models for symmetric relational data. *R package version*, 1, 2012.

[124] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. ROCR: Visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, October 2005. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bti623.

[125] Mark S Handcock, Garry Robins, Tom AB Snijders, Jim Moody, and Julian Besag. Assessing degeneracy in statistical models of social networks. Technical report, Citeseer, 2003.

[126] Pedro L Alonso, Graham Brown, Myriam Arevalo-Herrera, Fred Binka, Chetan Chitnis, Frank Collins, Ogobara K Doumbo, Brian Greenwood, B Fenton Hall, and Myron M Levine. A research agenda to underpin malaria eradication. *PLoS Med*, 8(1):e1000406, 2011. ISSN 1549-1676. 00409.

[127] Joel G Breman. Eradicating malaria. *Science progress*, 92(1):1–38, 2009. ISSN 0036-8504. 00050.

[128] The Centers for Population Health and Health Disparities Evaluation Working Group and Janet Okamoto. Scientific collaboration and team science: A social network analysis of the centers for population health and health disparities. *Translational Behavioral Medicine*, 5(1):12–23, March 2015. ISSN 1869-6716, 1613-9860. doi: 10.1007/s13142-014-0280-1. 00000.

[129] Eldon Y. Li, Chien Hsiang Liao, and Hsiuju Rebecca Yen. Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42 (9):1515–1530, November 2013. ISSN 00487333. doi: 10.1016/j.respol.2013.06.012. 00082.

[130] Michael E Bales, Stephen B Johnson, and Chunhua Weng. Social network analysis of interdisciplinarity in obesity research. volume 870, 2008. 00015.

[131] Michael E Bales, Stephen B Johnson, Jonathan W Keeling, Kathleen M Carley, Frank Kunkel, and Jacqueline A Merrill. Evolution of coauthorship in public health services and systems research. *American journal of preventive medicine*, 41 (1):112–117, 2011. ISSN 0749-3797. 00012.

[132] Caroline S. Wagner and Loet Leydesdorff. Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10): 1608–1618, December 2005. ISSN 00487333. doi: 10.1016/j.respol.2005.08.002. 00657.

[133] Omwoyo Bosire Onyancha and Jan Resenga Maluleka. Knowledge production through collaborative research in sub-Saharan Africa: How much do countries contribute to each other's knowledge output and citation impact? *Scientometrics*, 87(2):315–336, May 2011. ISSN 1588-2861. doi: 10.1007/s11192-010-0330-5.

[134] Christian S Schmid and Bruce A Desmarais. Exponential Random Graph Models with Big Networks: Maximum Pseudolikelihood Estimation and the Parametric Bootstrap. *arXiv preprint arXiv:1708.02598*, 2017.

[135] Thomas W. Valente, Kayo Fujimoto, Chih-Ping Chou, and Donna Spruijt-Metz. Adolescent Affiliations and Adiposity: A Social Network Analysis of Friendships and Obesity. *Journal of Adolescent Health*, 45(2):202–204, August 2009. ISSN 1054139X. doi: 10.1016/j.jadohealth.2009.01.007.

[136] Kayla de la Haye, Garry Robins, Philip Mohr, and Carlene Wilson. Obesity-related behaviors in adolescent friendship networks. *Social Networks*, 32 (3):161–167, July 2010. ISSN 03788733. doi: 10.1016/j.socnet.2009.09.001.

[137] Olga Kornienko, Katherine H Clemans, Dorothée Out, and Douglas A Granger. Hormones, behavior, and social network analysis: Exploring associations between cortisol, testosterone, and network structure. *Hormones and behavior*, 66(3): 534–544, 2014. ISSN 0018-506X.

[138] David R Hunter, Steven M Goodreau, and Mark S Handcock. Goodness of Fit of Social Network Models. *Journal of the American Statistical Association*, 103(481):

248–258, March 2008. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214507000000446.

[139] Luka Kronegger, Franc Mali, Anuška Ferligoj, and Patrick Doreian. Collaboration structures in Slovenian scientific communities. *Scientometrics*, 90(2):631–647, February 2012. ISSN 1588-2861. doi: 10.1007/s11192-011-0493-8.

[140] Caroline S Wagner, Irene Brahmakulam, Brian Jackson, Anny Wong, and Tatsuro Yoda. Science and technology collaboration: Building capability in developing countries. Technical report, RAND CORP SANTA MONICA CA, 2001.

[141] World Health Organization. Atlas of African Health Statistics 2016: Health situation analysis of the African Region. 2016. ISSN 9290232919.

[142] Zdenek Horak, Milos Kudelka, Vaclav Snasel, Ajith Abraham, and Hana Rezankova. Forcoa.NET: An interactive tool for exploring the significance of authorship networks in DBLP data. pages 261–266. IEEE, October 2011. ISBN 978-1-4577-1133-6 978-1-4577-1132-9 978-1-4577-1131-2. doi: 10.1109/CASON.2011.6085955.

[143] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

[144] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. Shiny: Web Application Framework for R. R package version 1.0. 3. 2017. URL `https://CRAN.R-project.org/package=shiny`.

[145] W Chang and Barbara Borges Ribeiro. Shinydashboard: Create Dashboards with 'Shiny'. *R package version 0.5*, 1, 2015.

[146] Jim Wilson. *Node. Js 8 the Right Way: Practical, Server-Side Javascript That Scales*. Pragmatic Bookshelf, 2018. ISBN 1-68050-536-X.

[147] Mark Newman. The physics of networks. *Physics today*, 61(11):33–38, 2008. ISSN 0031-9228.

[148] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, 22(2):338–342, September 2007. ISSN 0892-6638, 1530-6860. doi: 10.1096/fj.07-9492LSF. 00997.

[149] M.-E. Lynall, D. S. Bassett, R. Kerwin, P. J. McKenna, M. Kitzbichler, U. Muller, and E. Bullmore. Functional Connectivity and Brain Networks in Schizophrenia. *Journal of Neuroscience*, 30(28):9477–9487, July 2010. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.0333-10.2010.

[150] C. Grefkes and G. R. Fink. Reorganization of cerebral networks after stroke: New insights from neuroimaging with connectivity approaches. *Brain*, 134(5): 1264–1276, May 2011. ISSN 0006-8950, 1460-2156. doi: 10.1093/brain/awr033.

[151] Betty M. Tijms, Alle Meije Wink, Willem de Haan, Wiesje M. van der Flier, Cornelis J. Stam, Philip Scheltens, and Frederik Barkhof. Alzheimer's disease: Connecting findings from graph theoretical studies of brain networks.

Neurobiology of Aging, 34(8):2023–2036, August 2013. ISSN 01974580. doi: 10.1016/j.neurobiolaging.2013.02.020.

[152] Sean L. Simpson, Satoru Hayasaka, and Paul J. Laurienti. Exponential Random Graph Modeling for Complex Brain Networks. *PLoS ONE*, 6(5):e20039, May 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0020039.

[153] Catalina Obando Forero and Fabrizio De Vico Fallani. A statistical model for brain networks inferred from large-scale electrophysiological signals. 2017.

[154] Catalina Obando Forero and Fabrizio De Vico Fallani. Graph Models of Brain Connectivity Networks. 2015.

[155] F. De Vico Fallani, J. Richiardi, M. Chavez, and S. Achard. Graph analysis of functional brain networks: Practical issues in translational neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1653): 20130521–20130521, September 2014. ISSN 0962-8436, 1471-2970. doi: 10.1098/rstb.2013.0521.

[156] Peng Wang, Philippa Pattison, and Garry Robins. Exponential random graph model specifications for bipartite networks—A dependence hierarchy. *Social Networks*, 35(2):211–222, May 2013. ISSN 03788733. doi: 10.1016/j.socnet.2011.12.004.

[157] Michel R.T. Sinke, Rick M. Dijkhuizen, Alberto Caimo, Cornelis J. Stam, and Willem M. Otte. Bayesian exponential random graph modeling of whole-brain

structural networks across lifespan. *NeuroImage*, 135:79–91, July 2016. ISSN 10538119. doi: 10.1016/j.neuroimage.2016.04.066.

Appendix

Combined MEG and fMRI Exponential Random Graph Modeling for inferring functional Brain Connectivity.

Combined MEG and fMRI Exponential Random Graph Modeling for inferring functional Brain Connectivity.

Roseric Azondekon[1], Zachary James Harper[12], and Charles Michael Welzig[2]

[1]University of Wisconsin Milwaukee, Milwaukee, WI, USA

[2]Medical College of Wisconsin, Milwaukee, WI, USA

Abstract

Estimated connectomes by the means of neuroimaging techniques have enriched our knowledge of the organizational properties of the brain leading to the development of network-based clinical diagnostics. Unfortunately, to date, many of those network-based clinical diagnostics tools, based on the mere description of isolated instances of observed connectomes are noisy estimates of the true connectivity network. Modeling brain connectivity networks is therefore important to better explain the functional organization of the brain and allow inference of specific brain properties. In this report, we present pilot results on the modeling of combined MEG and fMRI neuroimaging data acquired during an n-back memory task experiment. We adopted a pooled Exponential Random Graph Model (ERGM) as a network statistical model to capture the underlying process in functional brain networks of 9 subjects' MEG and fMRI data out of 32 during a 0-back vs 2-back memory task experiment. Our results suggested strong evidence that all the functional connectomes of the 9 subjects have small world properties. A group level comparison using a non-parametric paired permutation t-test comparing the conditions pairwise showed no significant difference in the functional connectomes across the subjects. Our pooled ERGMs successfully reproduced important brain properties such as functional segregation and functional integration. However, the ERGMs reproducing the functional segregation of the brain networks discriminated between the 0-back and 2-back conditions while the models reproducing both properties failed to successfully discriminate between both conditions. The pilot results presented here are promising and would improve in robustness with a larger sample size. Nevertheless, our pilot results tend to support previous findings that functional segregation and integration are sufficient to statistically reproduce the main properties of brain network.

Keywords: Functional brain connectomes, ERGM, Functional connectivity, Neuroimaging data

BACKGROUND

The development of sophisticated neuroimaging techniques has enabled the acquisition of non-invasive quantitative data prompting to the development of new concept of the analyses of these data. In the existing literature, estimated connectomes by the means of neuroimaging techniques have enriched our knowledge of the organizational properties of the brain and enabled the development of network-based clinical diagnostics of certain pathologies such as schizophrenia [1], stroke [2], and Alzheimer's disease [3]. Although the mere descriptive analyses of the functional brain connectivity used in those researches have improved our knowledge of brain connectivity maps, there remains a gap in the literature since the description of isolated instances of observed connectivity network are noisy estimates of the true connectivity network [4,5]. In fact, the brain functioning can be represented as a connectivity network (connectome) where parcels or anatomical regions or regions of interest (ROIs) of the brain represent the vertices and the edges determine statistical dependency of combined neuronal activities between the vertices [6].

Modeling brain connectivity networks is therefore important to better explain the functional organization of the brain and to allow inference of specific brain properties. At first, three main mathematical models referred as null models or generative models have been proposed to infer some observed basic network properties such as network size, connection density, and degree distribution. The first is the simple random network model proposed by Erdős and Rényi [7]; a more general formulation of this model was described by Gilbert [8]. Random network models help to hypothesis testing whether the topology of a brain connectivity network arise purely by chance. The second model was proposed by Watts-Strogatz [9] and termed as the Watts-Strogatz small-world model. This model generates random networks spanning at the middle ground of the topological spectrum of random networks and lattice networks. Small world networks are characterized by a relatively high clustering coefficient and a small average path length between nodes. The third model is the preferential attachment model proposed by Barabási and Albert [10]. This model generates more realistic, scale-free degree distribution networks from the concept of "the rich get richer". Although these models allow hypothesis testing and the identification of relevant network properties, they come up short at explaining the organizational mechanisms of brain connectivity network formation [5]. In addition, these mathematical models are not estimable

from the observed data, do not allow fitness to the data, and hence cannot provide a reasonable representation of the observed network [11].

To remedy the limitations of generative null models, statistical models have been proposed to not only support inference but to capture and explain the process underlying the formation of the network structure. Unlike mathematical network models, statistical network models are designed to consider all the alternative networks estimated and weighted from observed data [12]. Furthermore, they specifically allow the assessment of significance of terms in the model and evaluation thanks to the goodness of fit. To date, three classes of such statistical network models have been proposed: the class of exponential random graph models, the class of stochastic block models, and the class of latent network models [11]. Analogous to standard regression models, the class of exponential random graph models (ERGM) also referred to as p* models (ERGM family models) appears as a flexible choice to simultaneously assess the role of specific network features in the overall organization of the complexity of brain networks. ERGM based connectivity analyses can help simulate and discriminate normal and abnormal brain organization and functioning [13]. In the social science literature, p* models prove successful at studying complex network interactions [14–18].

In neuroscience, the application of p* family models is still limited as very few studies have proved to successfully use them to model neuroimaging data based connectomes. To the best of our knowledge, the first study of this kind was reported in 2011 by Simpsons et. al [19] who applied ERGM on connectomes derived from 10 fMRI data collected from 10 subjects. Another study conducted in 2016 was reported by Sinke et. al [20] who applied Bayesian ERGM on diffusion tensor imaging (DTI) collected from 382 healthy subjects. More recently, in 2018, Obando and De Vico Fallani [5] published the first study to model functional connectomes derived from EEG data collected on 108 subjects during eyes-open (EO) and eyes-closed (EC) resting-state conditions. While it is understandable that all those studies pioneered the use of p* family models on neuroimaging connectomes, the applicability of ERGM family models to other connectomes inferred from other neuroimaging data is yet to be proved. In this report, we described how we applied p* models to combined MEG and fMRI neuroimaging data acquired during a memory task experiment.

There also remains many methodological unanswered issues such as the connectivity metrics to derive network topology, the ERGM terms to include in the modeling process, as well as how ERGM must be fit to the subject's connectomes. Simpsons et. al [19] and Obando and De Vico Fallani [5] for instance, fit a single ERG model to each subject data. Such a methodological approach lacks robustness when for example, one seeks to estimate a single model that discriminates between EO and EC resting state conditions. ERGM family models have a lot of potentials, especially in providing a better and more robust alternative network-based diagnostic model to the descriptive network-based diagnostic methods of medical conditions [1–3,21,22]. We address the lack of robustness from the previous studies by taking a pooled ERGM approach combining functional connectomes across subjects for each condition.

To the best of our knowledge, this report is the first to ever describe the application of ERGM to combined MEG and fMRI data.

METHODS

Participants

Participants were 32 healthy, right-handed adults between the ages of 18 and 40 recruited from the community using local print and electronic media. Recruited participants were all English speakers with at least 12 years of education. No exclusion was made on the basis of race, ethnicity, or gender. Because of the MRI scans, all participants were assessed for contraindications to MRI scanning, such as implanted electronic devices or ferrous metal in sensitive areas.

Experiment

The participants were asked to perform n-back memory tasks during MEG scans. In our n-back tasks, participants are presented a sequence of visual stimuli one-by-one. For each stimulus, they need to decide if the current stimulus is the same as the one presented n trials ago. Specifically, the participants performed 0-back and 2-back memory tasks during which they are asked to match geometric shapes. The MEG paradigm consists of nine experimental blocks: two blocks each of matching pictures with five control blocks, each lasting 32s for a total scan length of 4:48 min.

Each block begins with a brief instruction statement: “Match Faces” or “Match Tools”. Each matching block consists of six images. For each face block, three images of each gender and target affect is presented. All images are presented sequentially, with no inter-stimulus interval, for a period of 5s and in a randomized fashion for both 0-back and 2-back memory tasks. The order of the paradigm is counterbalanced across subjects. During MEG recordings, subjects respond by pressing a button on one of two button boxes, allowing for the determination of accuracy and reaction time.

MEG acquisition

All participants undergo MEG scanning at the Medical College of Wisconsin (MCW) MEG lab. Before the experiment, a Polhemus Isotrak® system is used to digitize participants’ cardinal landmarks (nasion and pre-auricular points) and head shape. Four head position indicator coils are fixed to the participants’ head and referenced to the other digitized landmarks. Two electrodes are placed along the plane of the chest to collect ECG signal. MEG data are acquired with the participant seated upright in the scanner. Data are sampled at 2,000 Hz. The scanning session consists in two to five runs of 10 minutes each. Prior to each subject’s scanning session, one to two runs of five to 10 minutes each of empty room MEG data are recorded for noise characterization. In addition, one to two runs of 10 minutes of Eyes-Open (EO) resting state of MEG data are also recorded after the experimental runs. All MEG scanning sessions take place on a different day than MRI scanning sessions.

MRI acquisition

All participants undergo high-resolution T1-weighted structural MRI at the MCW 7 Tesla MRI facility. MRI scanning sessions include localizer scans and a GE SPGR T1 acquisition with approximately 1x1x1 mm voxel size and parameters optimized for grey-white contrast. For each subject, the scanning session requires approximately 90 minutes and takes place on a different day than the MEG scanning session.

Data processing

MRI data

The fMRI data are processed using FreeSurfer [23], thanks to which, the brain is anatomically parcellated into 68 Regions of Interest (ROIs) or parcels using the automatic parcellation ('aparc') annotation. A neuroanatomical label is assigned to each ROI on a cortical surface model based on probabilistic information estimated from a manually labeled training set [23,24].

MEG data

We apply MaxFilter, an essential pre-processing tool for MEG data, in order to remove noise sources likely to originate from outside the sensor array. We then transform the MEG data using the temporally extended signal space separation method (tSSS) to remove strong interference caused by external and nearby sources. The tSSS-reconstructed MEG data are processed using MNE-Python [25,26], an open source Python library for the processing of EEG and MEG data. Next, the data are cleaned using Independent Component Analysis (ICA) to remove EOG and ECG artifacts. For each subject, the MEG recordings are co-registered to the anatomical fMRI preprocessed data. BEM, source, and forward solution for each run are then computed. Next, the MEG data are resampled at 500Hz, and notch filtered at 60Hz. Further filtering including low and high band filters at respectively 50Hz and 1Hz are applied as well. For each subject, the recording MEG runs are further concatenated in one single raw file. The precomputed forward solutions are averaged across runs and a covariance matrix is computed from the empty room MEG runs. The forward solution and the covariance matrix are used to compute an inverse solution. Using detected event ids corresponding to the stimuli presentation, we next proceed to the extraction of the events. The extracted events are epoched accordingly. From the previously computed inverse solution, the inverse operator is determined and applied to each of the epoched 0-back and 2-back conditions separately. The resting state MEG runs are processed similarly to the experimental runs without the event detection step. For each 0-back, 2-back, and resting state conditions, we compute the spectral coherence [27] to measure functional connectivity (FC) between MEG signals of ROIs or parcels x and y at a specific frequency band f as follows:

$$SC_{xy}(f) = \frac{\left|S_{xy}(f)\right|^2}{S_{xx}(f)S_{yy}(f)}$$

where S_{xy} is the cross-spectrum between x and y, and S_{xx} and S_{yy} are respectively the autospectra of x and y. The connectivity matrix $SC(f)$ of size 68 × 68 where the entry $SC_{xy}(f)$ contains the value of the spectral coherence between the MEG signals of ROIs or parcels x and y at the frequency f. The connectivity matrices are computed at each and across *theta* (4 – 8Hz), *alpha* (8 – 15Hz), *beta* (15 – 35Hz), and *gamma* (35 – 120Hz) frequency bands. All data processing is performed using MNE-Python, an Open-source Python software [28].

Network generation

The computed connectivity matrices are adjacency symmetric matrices representing undirected weighted network, where the vertices are the 68 ROIs or brain parcels generated from the 'aparc' annotation and the edges are weighted by the magnitude of the spectral coherence. The adjacency matrices are then filtered to obtain the strongest edges in each brain network. While various studies [6,29–31] recommend different filtering techniques of the adjacency matrix, we decide to set an arbitrary threshold depending on each connectivity matrix. Using NetworkX [32], a python library for exploring complex networks, we generate binary functional brain connectivity networks from the filtered adjacency matrices. Each one of the graphs are exported in a graphml format for model estimation in R, an open-source environment for statistical computing [33].

Assessing the small worldness of the connectivity networks

Small world networks interposed between random and lattice networks. Like a regular lattice, they show high clustering and like regular random networks, they display low average path length. While the high clustering supports degeneracy and triangular integration, and may facilitate functional specialization, the low average path length facilitates efficient integration across the brain network. Since healthy brain networks have been proved to have small world organization [9], these two properties of small world networks have been used in clinical applications, particularly in the classification of brain disorders. [34,35]. To assess the small worldness of the generated functional connectivity networks, there remains the question regarding which clustering coefficient values should be considered high and which average path length values should be deemed as low. To address this question, Fornito et al. [36] propose a simple solution which

consists in comparing the clustering and average path length values in each of the observed functional connectivity networks to comparable values computed in appropriately randomized control networks. Consequently, two indices which we adopt here, are defined:

- The normalized clustering coefficient γ defined as:

$$\gamma = \frac{Cl}{\langle Cl_{rand} \rangle}$$

Where Cl_{rand} is the average clustering coefficient computed over an ensemble of randomized surrogate network and Cl is the average clustering coefficient of the observed network defined as:

$$Cl = \frac{1}{N} \sum_{i \epsilon N} \frac{2t_i}{k_i(k_i - 1)}$$

Where N is the number of nodes, k_i is the degree of node i, and t_i is the number of closed triangles attached to node i in the observed network.

- The normalized measure of path length λ defined as:

$$\lambda = \frac{L}{\langle L_{rand} \rangle}$$

Where $\langle L_{rand} \rangle$ is the mean of the average path length computed over an ensemble of randomized surrogate network, and L is the observed average path length defined as:

$$L = \frac{1}{N(N-1)} \sum_{i,j \epsilon N; i \neq j} d_{ij}$$

Where d_{ij} is the distance of the shortest path, between nodes i and j.

In a small world network therefore, one would expect $\lambda \sim 1$ and $\gamma > 1$.

Humphries et al. [37] proposed the ratio of γ and λ as a single scalar index to quantify the small-worldness of a network:

$$\sigma = \frac{\gamma}{\lambda}$$

A network with small world properties should be associated with a value of σ greater than 1.

For each of the connectivity networks, we constructed an ensemble of 1,000 surrogate random networks using Monte-Carlo based simulations. We next compute respectively γ, λ, and σ as defined above. Any network with a value of σ greater than 1 is characterized as having small world properties. Since all our data have been recorded from "healthy individuals", we expect all the functional connectivity networks to display small world organization across all three conditions (0-back vs 2-back vs resting state).

Statistical Group Analysis

After the computation of the spectral connectivity in MNE-Python, the ROIs are exported in MNI coordinates in millimeters. The connectivity matrices are also exported as connectivity matrix files. Each matrix file contains the 68 lines by 68 columns of connectivity values. We then use the Network Based Statistic Toolbox (NBS) developed in Matlab by Zalesky et al. [38] to compare the brain networks between conditions. We used a non-parametric paired permutation t-test comparing the three conditions pairwise with a statistical significance level set at 0.05. The number of permutations is set at 100,000 for each comparison.

Exponential Random Graph Model Estimation

Given a network graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, let the matrix $\mathbf{Y} = [Y_{ij}]$, be the random adjacency matrix of G. Each entry Y_{ij} denotes a binary variable indicating the presence or absence of edge between two vertices i and j. Since our brain connectivity network is an undirected network, $Y_{ij} = Y_{ji}$. Let's denote the matrix $\mathbf{y} = [y_{ij}]$ a particular realization of $\mathbf{Y}$. The general formulation of ERGM has the form [11]:

$$\mathbb{P}_\theta(\mathbf{Y} = \mathbf{y}) = \left(\frac{1}{\kappa(\theta)}\right) exp\left\{\sum_H \theta_H g_H(\mathbf{y})\right\}$$

Where H is a configuration in G, $g_H(\mathbf{y}) = \prod_{y_{ij} \epsilon H} y_{ij}$, θ is a vector of parameter, and $\kappa(\theta)$ is a normalization constant defined as:

$$\kappa(\theta) = \sum_{y} exp\left\{\sum_{H} \theta_H g_H(\mathbf{y})\right\}$$

Several variants of ERGM have been proposed [39], here we rely on the temporal ERGM variant proposed by Leifeld et al. [40] which applied without any temporal dependencies corresponds to a pooled ERGM. We refer the reader to Leifeld et al. [40] for a detailed explanation of the model. Our main assumption justifying this choice is that different brain processes are involved in the 0-back, and 2-back memory tasks. Therefore, all changes in the functional connectivity brain networks under each condition are attributable to variation according to an underlying ERGM. Since the subjects are dependent from each other, the estimates of the pooled ERGM reflect the average effects across all the subjects' brain networks under a specific condition.

We model several organizational and functional mechanisms of the brain including functional segregation and functional integration [41,42]. Functional integration refers to distributed processes defining brain function and is measured in connectomics by the average path length (already defined above) or the global efficiency E_g defined as:

$$E_g = \frac{1}{N(N-1)} \sum_{i,j \epsilon N; i \neq j} \frac{1}{d_{ij}}$$

Functional segregation refers to the idea that all vertices in the brain network (or ROIs or brain parcels) will display divergent pattern of activity and hence be statistically independent. In connectomics, functional segregation is measured by the clustering coefficient (already defined above) and the local efficiency E_l defined as:

$$E_l = \frac{1}{N} \sum_{i \epsilon N} E_g(G_i)$$

Where G_i is the subgraph formed by the vertices connected to i.

Model construction and estimation are computed using the statistical software R [33]. In the **btergm** R package that we used, functional integration and functional segregation are already respectively coded as the GWNSP (Geometrically Weighted Nonedgewise Shared Partner distribution) and the GWDSP (Geometrically Weighted Dyadwise Shared Partner distribution) ERGM terms [6]. We also model other ERGM terms including degree distribution, k-triangles (for

transitivity) and k-stars (for highly connected vertices). We assess the Goodness-Of-Fit (GOF) of each model, simulating 1,000 networks from the estimated model and comparing them to the observed networks. The best model is selected based on the lowest Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) and the highest log likelihood.

The R packages **igraph** [43], **sna** [44] and **network** [45] are also used for the manipulation of the brain network graphs. All computations are performed in Rstudio-server setup on a 64 cores CPU server equipped with a 512GB RAM.

RESULTS

In this section, we present pilot results based on a subset of nine subjects out of the 32 participants we collected neuroimaging data from. Likewise, only results on the 0-back and 2-back conditions are presented as the resting state data were yet to be processed at the pilot stage. Also, these pilot results were obtained from the connectivity matrices computed across all the frequency bands.

Small-worldness Assessment

The results of the small-worldness assessment are presented in table 1. As we can see, across all subjects for the 0-back and the 2-back conditions, the functional connectivity brain network have values for γ that are larger than one and values for α that are close to one. Consequently, the values for σ are all larger than one. This is a strong evidence suggesting that all the functional connectivity brain networks have small-world properties.

Table 1. Small-worldness assessment of the brain networks based on 1000 randomized control surrogates

	0-back			2-back		
subjects	γ	λ	σ	γ	λ	σ
1	4	1.107	3.613	4	1.104	3.623
2	3	1.156	2.595	3	1.157	2.593
3	4	1.167	3.428	4	1.226	3.263
4	5	1.16	4.31	5	1.161	4.307
5	4	1.142	3.503	4	1.18	3.39
6	4	1.214	3.295	4	1.205	3.32
7	3	1.113	2.695	3	1.12	2.679

Group level comparison

In the group level analysis, we see no significant difference ($p>>0.1$) between the connectivity values of the 0-back versus the 2-back conditions across all and for each of the frequency bands. This lack of significance is illustrated in figure 1 which displays the 300 strongest connections between the identified ROIs or brain parcels.

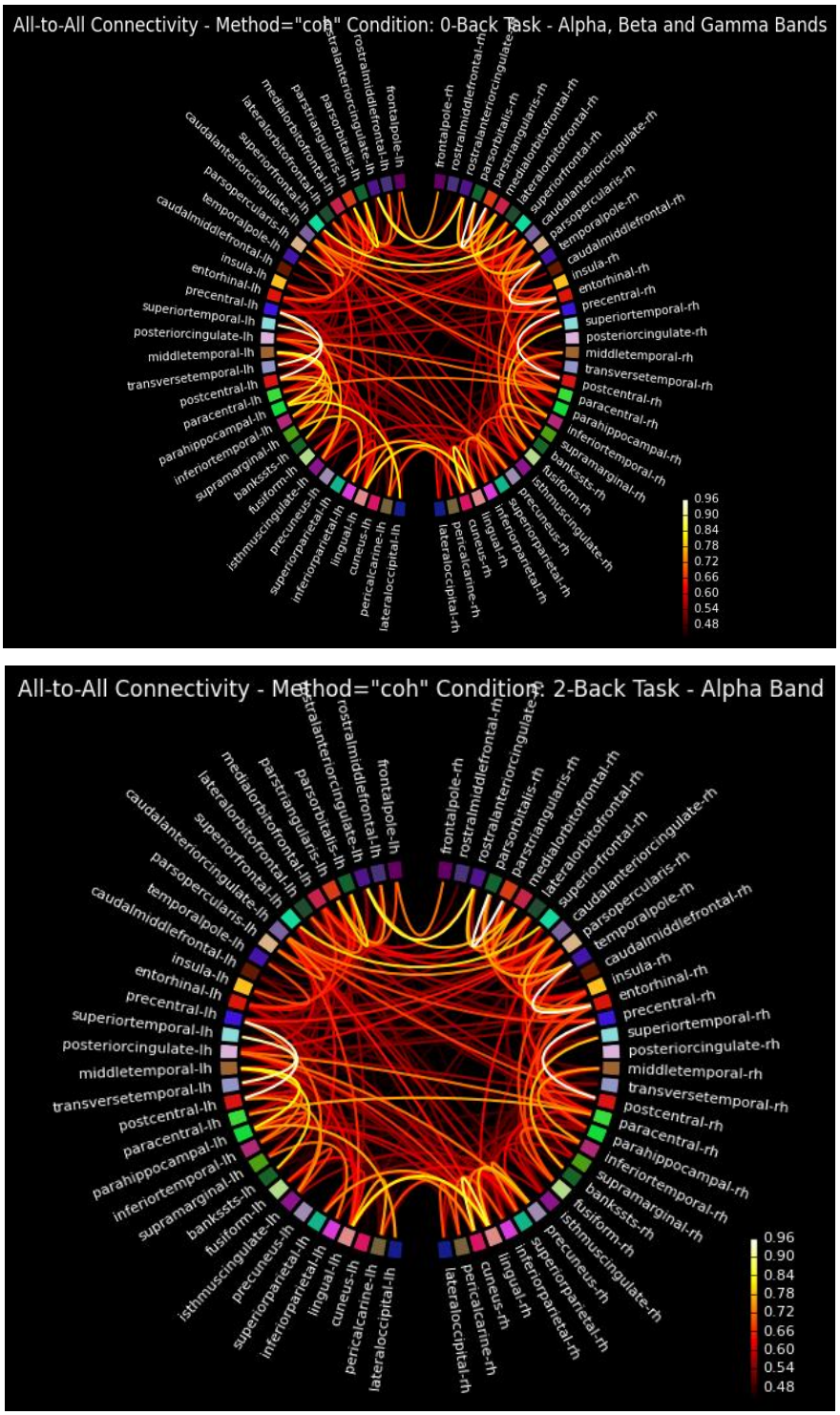


Figure 1. Connectivity plots of the 0-back (top) compared to the 2-back (bottom) memory tasks in subject 1.

Exponential Random Graph Model

Most of the model configurations we fit did not converge and/or degenerate. Table 2 presents the configurations of ERGMs we successfully fit at this pilot stage. At this stage, none of our model configurations containing the k-star or triangle ERGM terms was successful. All of them degenerated around 50 iterations and did not converge.

Table 2. Successful ERG model configurations

Models	**edges**	**degree**	**GWDSP**	**GWNSP**
Null	x			
Model 1			x	
Model 2	x	x	x	
Model 3	x	x	x	x

Table 3 presents the estimates of the ERGM configurations. We can see that the model estimates were all significantly higher than zero. However, for the null and model3, the confidence intervals of the estimates for the 0-back and 2-back conditions overlap meaning that those models failed to discriminate between both conditions. On the other hand, model1 and model2 discriminate between 0-back and 2-back conditions as the model estimates were significantly different than zero and their confidence intervals do not overlap. Overall, the ERGM model containing both the functional segregation and functional integration did not prove successful at discriminating between the 0-back and the 2-back conditions (see coefficient plot at Figure 2).

Given the low sample size, the AIC, BIC, and the log-likelihood were only computed for the null model. We could not efficiently compare the models according to those values. However, model3 containing the functional segregation and functional integration ERGM terms proves interesting as we believe an increase in the sample size would tremendously improve it at discriminating between 0-back and 2-back conditions.

Figure 4 shows the GOF plot of model 3 for both 0-back and 2-back conditions. The plain black line represents the feature distribution from the observed brain networks and the dashed black line is the feature distribution from the 1,000 simulated networks from model3. We expect both lines to overlap when the model captures the underlying ERGM process. As we can see in the GOF in

figure 4, model3 captures well the underlying ERGM process for 0-back and 2-back. However, the simulated walktrap modularity distribution (in red) does not match well the observed one (in black).

Conclusion

In this report, we use a pooled variant of ERGM to capture differentially the underlying ERGM process involved in two nback memory tasks. Our models perform decently well given the significant model estimates. The low sample size of the brain networks is a tangible reason justifying the failure of most of the model configurations we attempted to fit. Consequently, we could not compare the model according to the AIC, BIC and the log-likelihood values. A larger sample size would enable a better model specification. The insignificant difference between the connectivity values of the 0-back and 2-back conditions at the group level comparison has been confirmed at the statistical modeling step. Nevertheless, the pilot results presented here are promising and would improve in robustness when all the remaining pre-processing will be completed and integrated to the analysis. While our results are not complete, they tend to support previous findings reported by De Vico Fallani et al. [5] that functional segregation and integration are sufficient to statistically reproduce the main properties of brain network.

It is worth noting that our connectivity networks were computed across all the frequency bands. Also, it would have been interesting to compare the resting state connectivity network pairwise with the ones of the 0-back and 2-back conditions. Unfortunately, those data were not pre-processed enough to be included in the analyses.

Finally, the connectivity values in this report are estimated by means of the spectral coherence which is known to suffer from possible volume conduction effects [46]. Other measures of connectivity such as Phase Lag Index (PLI), Phase-Locking Value (PLV), coherency, or the Imaginary coherence are potential alternatives worth considering. Although a binarizing threshold may influence the topology of the network, our thresholding procedure to filter the connectivity value and binarize the strongest edges has been based on the observation of the connectivity plot. A density based thresholding procedure has been proposed in [5] and proved to ensure a meaningful network.

Table 3.

	0back				2back			
	Null	**Model1**	**Model2**	**Model3**	**Null**	**Model1**	**Model2**	**Model3**
edges	-2.05 [-2.09; -2.00]*		-0.85 [-0.94; -0.76]*	-2.63 [-2.83; -2.43]*	-2.06 [-2.10; -2.02]*		1.26 [1.21; 1.26]***	-2.73 [-2.91; -2.55]*
degree			1.64 [1.10; 2.18]*	0.59 [0.09; 1.10]*			-4.22 [-4.39; -4.05]***	0.26 [-0.24; 0.75]
GWDSP		-0.34 [-0.34; -0.33]*	-0.22 [-0.24; -0.20]*	1.48 [1.29; 1.67]*		-0.54 [-0.57; -0.52]*	-41.16	1.47 [1.31; 1.64]*
GWNSP				-1.72 [-1.92; -1.53]*				-1.71 [-1.87; -1.55]*
AIC	145963.24				245249.42			
BIC	145982.51				245269.7			
Log Likelihood	-72979.62				-122622.71			

***p < 0.001, **p < 0.01, *p < 0.05 (or 0 outside the confidence interval)

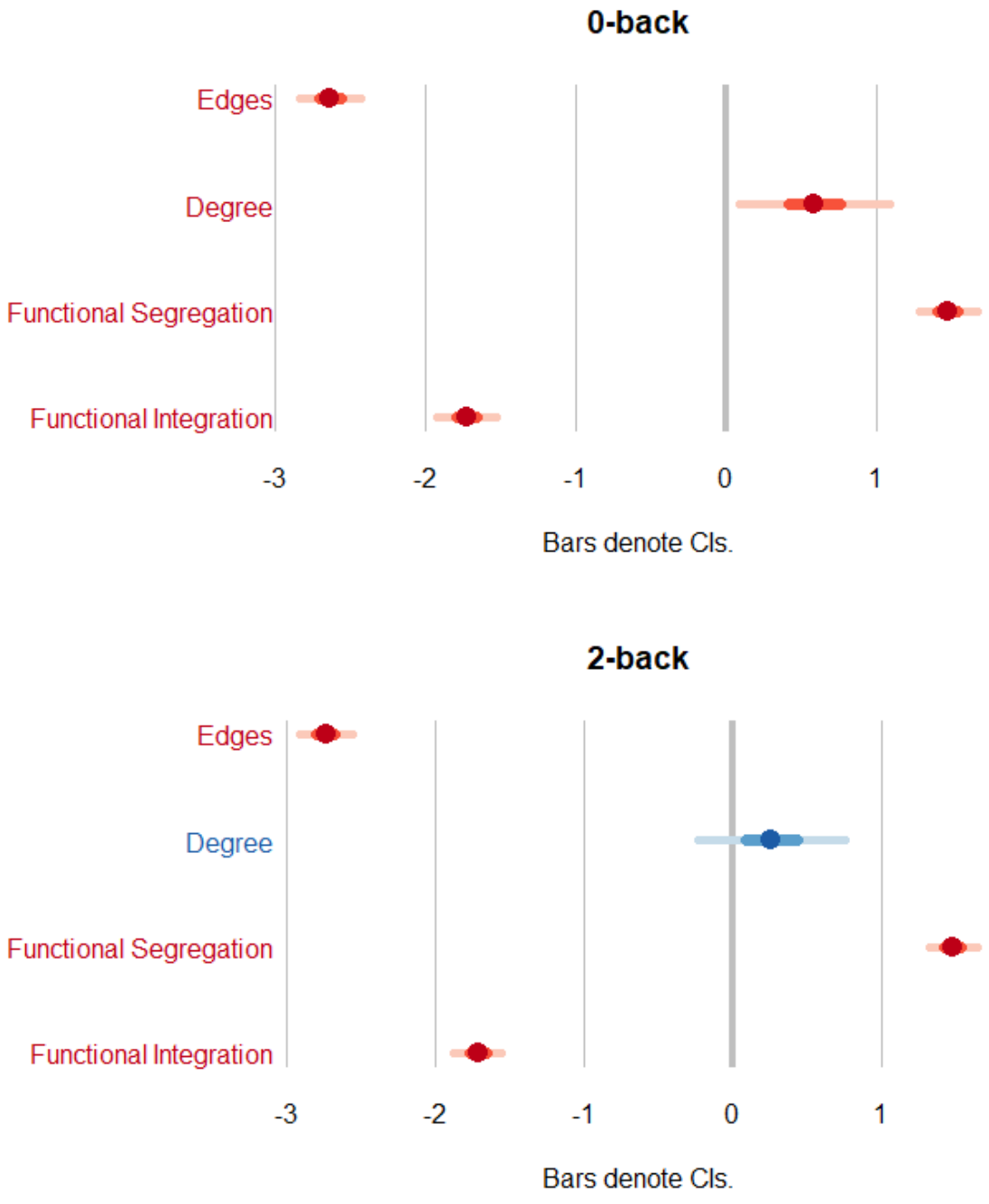
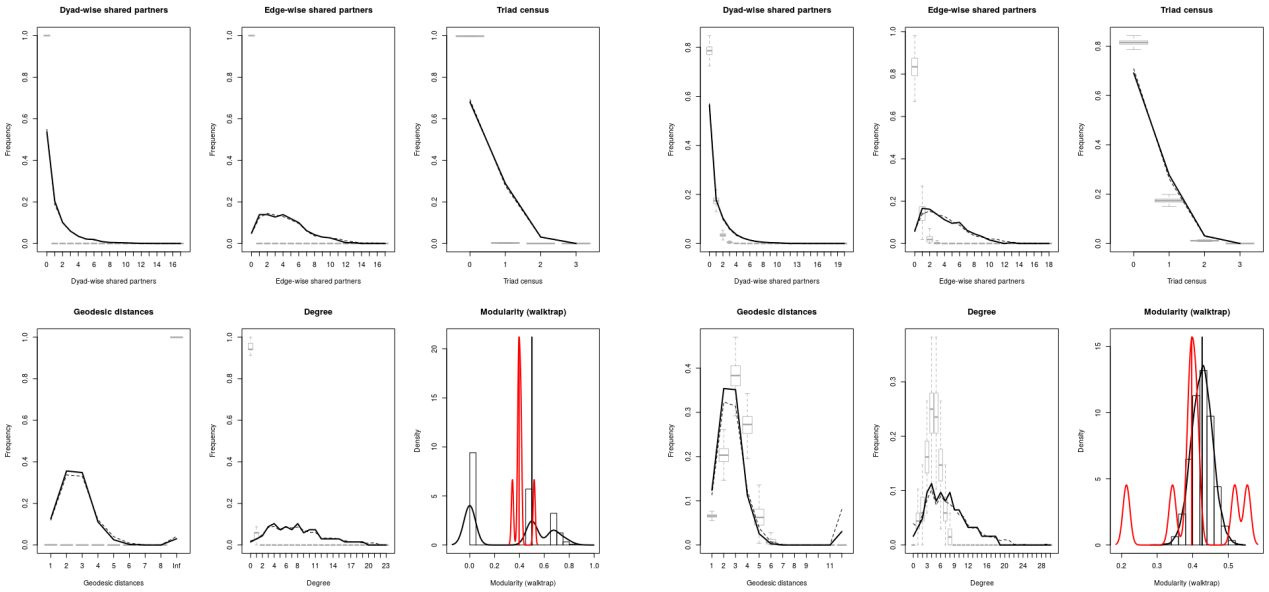


Figure 2. Coefficient plot of model3 comparing 0-back and 2-back conditions



(a) Model3 GOF for 0-back condition

(b) Model3 GOF for 2-back condition

Figure 3. GOF of model3 comparing 0-back and 2-back conditions.

Reference

1. Lynall M-E, Bassett DS, Kerwin R, McKenna PJ, Kitzbichler M, Muller U, Bullmore E. Functional Connectivity and Brain Networks in Schizophrenia. J Neurosci. 2010 Jul 14;30(28):9477–87.

2. Grefkes C, Fink GR. Reorganization of cerebral networks after stroke: new insights from neuroimaging with connectivity approaches. Brain. 2011 May 1;134(5):1264–76.

3. Tijms BM, Wink AM, de Haan W, van der Flier WM, Stam CJ, Scheltens P, Barkhof F. Alzheimer's disease: connecting findings from graph theoretical studies of brain networks. Neurobiol Aging. 2013 Aug;34(8):2023–36.

4. De Vico Fallani F, Richiardi J, Chavez M, Achard S. Graph analysis of functional brain networks: practical issues in translational neuroscience. Philos Trans R Soc B Biol Sci. 2014 Sep 1;369(1653):20130521–20130521.

5. Obando C, De Vico Fallani F. A statistical model for brain networks inferred from large-scale electrophysiological signals. J R Soc Interface. 2017 Mar;14(128):20160940.

6. Forero CO, Fallani FDV. Graph Models of Brain Connectivity Networks. 2015;

7. Erdos P, Rényi A. On the evolution of random graphs. Publ Math Inst Hung Acad Sci. 1960;5(1):17–60.

8. Gilbert EN. Random graphs. Ann Math Stat. 1959;30(4):1141–4.

9. Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. nature. 1998;393(6684):440–2.

10. Barabási A-L, Albert R. Emergence of scaling in random networks. science. 1999;286(5439):509–12.

11. Kolaczyk ED, Csárdi G. Statistical analysis of network data with R. Vol. 65. Springer; 2014.

12. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network Motifs: Simple Building Blocks of Complex Networks. Science. 2002 Oct 25;298(5594):824–7.

13. Fox MD, Greicius M. Clinical applications of resting state functional connectivity. Front Syst Neurosci. 2010;4:19.

14. Krivitsky PN, Goodreau SM. STERGM-Separable Temporal ERGMs for modeling discrete relational dynamics with statnet. 2017;

15. Goodreau SM, Kitts JA, Morris M. Birds of a Feather, Or Friend of a Friend?: Using Exponential Random Graph Models to Investigate Adolescent Social Networks. Demography. 2009;46(1):103–25.

16. Kornienko O, Clemans KH, Out D, Granger DA. Hormones, behavior, and social network analysis: Exploring associations between cortisol, testosterone, and network structure. Horm Behav. 2014;66(3):534–44.

17. Wang P, Pattison P, Robins G. Exponential random graph model specifications for bipartite networks—A dependence hierarchy. Soc Netw. 2013 May;35(2):211–22.

18. Niekamp A-M, Mercken LAG, Hoebe CJPA, Dukers-Muijrers NHTM. A sexual affiliation network of swingers, heterosexuals practicing risk behaviours that potentiate the spread of sexually transmitted infections: A two-mode approach. Soc Netw. 2013 May;35(2):223–36.

19. Simpson SL, Hayasaka S, Laurienti PJ. Exponential Random Graph Modeling for Complex Brain Networks. Sporns O, editor. PLoS ONE. 2011 May 25;6(5):e20039.

20. Sinke MRT, Dijkhuizen RM, Caimo A, Stam CJ, Otte WM. Bayesian exponential random graph modeling of whole-brain structural networks across lifespan. NeuroImage. 2016 Jul;135:79–91.

21. Achard S, Delon-Martin C, Vertes PE, Renard F, Schenck M, Schneider F, Heinrich C, Kremer S, Bullmore ET. Hubs of brain functional networks are radically reorganized in comatose patients. Proc Natl Acad Sci. 2012 Dec 11;109(50):20608–13.

22. Chennu S, Finoia P, Kamau E, Allanson J, Williams GB, Monti MM, Noreika V, Arnatkeviciute A, Canales-Johnson A, Olivares F, Cabezas-Soto D, Menon DK, Pickard JD, Owen AM, Bekinschtein TA. Spectral Signatures of Reorganised Brain Networks in Disorders of Consciousness. Ermentrout B, editor. PLoS Comput Biol. 2014 Oct 16;10(10):e1003887.

23. Fischl B. FreeSurfer. Neuroimage. 2012;62(2):774–81.

24. FreeSurfer I, FSL's F. FreeSurfer Tutorial.

25. Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS. MNE software for processing MEG and EEG data. Neuroimage. 2014;86:446–60.

26. Dinh C, Luessi M, Sun L, Haueisen J, Hamalainen MS. MNE-X: MEG/EEG Real-time acquisition, real-time processing, and real-time source localization framework. Biomed Eng Tech. 2013;

27. Carter GC. Coherence and time delay estimation. Proc IEEE. 1987;75(2):236–55.

28. Gramfort A. MEG and EEG data analysis with MNE-Python. Front Neurosci [Internet]. 2013 [cited 2018 Apr 30];7. Available from: http://journal.frontiersin.org/article/10.3389/fnins.2013.00267/abstract

29. Rubinov M, Sporns O. Complex network measures of brain connectivity: uses and interpretations. Neuroimage. 2010;52(3):1059–69.

30. Rubinov M, Sporns O. Weight-conserving characterization of complex functional brain networks. Neuroimage. 2011;56(4):2068–79.

31. Ginestet CE, Nichols TE, Bullmore ET, Simmons A. Brain network analysis: separating cost from topology using cost-integration. PloS One. 2011;6(7):e21570.

32. Hagberg A, Schult D, Swart P, Conway D, Séguin-Charbonneau L, Ellison C, Edwards B, Torrents J. Networkx. High productivity software for complex networks. Webová Strá Nka Httpsnetworkx Lanl Govwiki. 2013;

33. Team RC. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013. 2014;

34. Sporns O. The Non-Random Brain: Efficiency, Economy, and Complex Dynamics. Front Comput Neurosci [Internet]. 2011 [cited 2018 Apr 30];5. Available from: http://journal.frontiersin.org/article/10.3389/fncom.2011.00005/abstract

35. Stam CJ. Modern network science of neurological disorders. Nat Rev Neurosci. 2014 Sep 4;15(10):683–95.

36. Fornito A, Zalesky A, Bullmore ET. Fundamentals of brain network analysis. 2016.

37. Humphries M., Gurney K, Prescott T. The brainstem reticular formation is a small-world, not scale-free, network. Proc R Soc B Biol Sci. 2006 Feb 22;273(1585):503–11.

38. Zalesky A, Fornito A, Bullmore ET. Network-based statistic: Identifying differences in brain networks. NeuroImage. 2010 Dec;53(4):1197–207.

39. Wasserman S, Pattison P. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. Psychometrika. 1996 Sep;61(3):401–25.

40. Leifeld P, Cranmer SJ, Desmarais BA. Temporal Exponential Random Graph Models with xergm: Estimation and Bootstrap Confidence Intervals. J Stat Softw. 2015;

41. Tononi G, Edelman GM, Sporns O. Complexity and coherency: integrating information in the brain. Trends Cogn Sci. 1998;2(12):474–84.

42. Zeki S, Shipp S. The functional logic of cortical connections. Nature. 1988;335(6188):311.

43. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal Complex Syst. 2006;1695(5):1–9.

44. Butts CT. sna: Tools for Social Network Analysis. R package version 2.2-0. 2010;

45. Handcock MS, Hunter DR, Butts CT, Goodreau SM, Morris M. statnet: Software tools for the representation, visualization, analysis and simulation of network data. J Stat Softw. 2008;24(1):1548.

46. Srinivasan R, Winter WR, Ding J, Nunez PL. EEG and MEG coherence: Measures of functional connectivity at distinct spatial scales of neocortical dynamics. J Neurosci Methods. 2007 Oct;166(1):41–52.

ROSERIC AZONDEKON

roseric_2000@yahoo.fr
https://www.researchgate.net/profile/Roseric_Azondekon

Education

Sep 2014– August 2018 **PhD in Biomedical and Health Informatics (Expected graduation: Summer 2018)**
College of Engineering and Applied Science
University of Wisconsin Milwaukee
Milwaukee, United States

Sep 2012 – May 2014 **Master of Public Health, Epidemiology**
School of Public Health and Health Science
University of Massachusetts Amherst
Amherst, MA, United States

Dec 2007 – Mar 2010 **Master of Science, Applied Entomology**
Faculty of Science and Technology
University of Abomey-Calavi
Abomey-Calavi, Benin

Oct 2003 – Sep 2007 **Engineer degree, Biomedical Sciences**
Polytechnic School of Abomey-Calavi
University of Abomey-Calavi
Abomey-Calavi, Benin

Work Experience

Sep 2016 – April 2018 **Clinical Research Assistant I**
Medical College of Wisconsin, Welzig Computational Neuroscience lab, 8701 Watertown Plank Road, TBRC C2020, Tel: +1-414-955-5752, Email: welzig@mcw.edu
Milwaukee, WI 53226, Wisconsin, United States

Main activities Technical skills and responsibilities:

- Development of predictive models using Machine Learning techniques such as Deep Learning
- High Throughput Computing and Parallel Programming

- Acquisition, Processing and Analysis of MEG, EEG, fNIRS and fMRI data

sSep 2015 – Dec 2015

Teaching Assistant
University of Wisconsin - Milwaukee, Joseph J. Zilber School of Public Health, 1240 N 10th Street, Tel: +1-414-227-5006, Milwaukee, WI 53205, USA

Main activities Technical skills and responsibilities:

- Teaching Elementary data analysis using SAS to first-year MPH students enrolled in the Introductory Biostatistics course PH 702 801-LAB 14B (SAS Lab)
- Design of quizzes and various assignments to evaluate students' progress
- Provide grading and feedback to students

Nov 2014 – May 2015

Research Assistant
University of Wisconsin - Milwaukee, Laboratory for Public Health Informatics and Genomics1240 N 10th Street, Tel: +1-414-227-5006, Milwaukee, WI 53205, USA

Main activities Technical skills and responsibilities:

- Genomics data analysis of RNA-seq data using R, TCGA tools, Galaxy and Chipster
- Genomics pipeline development of various Bioinformatics tools such as VirusSeq
- High Throughput Computing
- Analysis of gene network

Dec 2006 – now (more than 11 years of experience in Medical Entomology)

Research Assistant, Head of the Quality Control of Insecticide Treated Tool Laboratory
Centre de Recherche Entomologique de Cotonou, PO-BOX : 06-2604, Tel.: +22921330825 – Email: akogbetom@yahoo.fr, Ministry of Health, Republic of Benin.

Main activities Technical skills and responsibilities:

- Monitoring of malaria vectors resistance to insecticide using WHO tube test and CDC bottle test and molecular analysis of mosquito species and resistance mechanisms such as Kdr, Ace-1R, Esterase, Oxydase and GST-Transferase
- Quality control of Indoor Residual Spray using wall cone Bioassay and filter paper for spray quality control with HPLC

- Efficacy assessment of Long Lasting Impregnated Nets in laboratory and field conditions using cone bioassays, Rapid Field Colorimetric Testing and Gas chromatography
- Identification and characterization of malaria vectors
- Monitoring and Evaluation of Long Lasting Impregnated Nets durability in field conditions (Net coverage, net survival and physical integrity)
- Mosquito larvae collection and rearing to adult stage in the insectarium
- Adult mosquito collection using Human Landing Catch, Indoor Pyrethrum Spray and Window traps
- Good knowledge of the mapping of malaria vectors insecticide resistance.

Specialization Certificates

Sep 2015 Programming for Everybody (Python), School of Information, University of Michigan Coursera Courses.

Feb 2009 Training workshop on the analysis of quantitative and qualitative data and an introduction to STATA and QSR Nvivo software, Bobo-Dioulasso, Burkina Faso supported by WHO/TDR/MIM Project ID no A 50066, from February 16th to 20th, 2009.

Computer Skills/Programming languages

Data analysis software High proficiency in R, SAS, STATA, Epilnfo, SPSS, Minitab, and QSR Nvivo

Mobile Data Collection ODK Collect and Makina Collect

Programming Languages High Proficiency in R, Python, SQL and Latex
Working knowledge of C, Javascript, HTML, and VBA

Awards & Grants

Jan 2016 Scholarship: IDB Merit Scholarship Programme for High Technology

Mar 2015 Fellowship: Distinguished Graduate Student Fellowship - University of Wisconsin Milwaukee

Aug 2014 Award: Chancellor's Graduate Student Award - University of Wisconsin Milwaukee

May 2013 Award: Hosmer BioStat Scholarship - University of Massachusetts Amherst

Jul 2012 Scholarship: **Fulbright Foreign Student Program**

Jun 2008 Scholarship: Islamic Development Bank M.Sc Scholarship Programme

Mar 2010	Award: Outstanding Student for the Master of Applied Medical Entomology Program at the University of Abomey-Calavi
Sep 2007	Award: Outstanding Student for the Engineer Degree in Biomedical Science Program at the Polytechnic School of the University of Abomey-Calavi
September 2003	Scholarship: Four-year Government scholarship awarded after a national competitive selection to enter the Biomedical Science program at the Polytechnic School of the University of Abomey-Calavi, Republic of Benin.

Skills & Activities

Skills	Monitoring & Evaluation, Mobile Data Collection, Data Management, Epidemiological Analysis, Vector Biology and Control, Infectious Disease Surveillance, Machine Learning, Natural Language Processing, Data Mining, Complex Analysis of Network Data
Proficient Languages	English, French
Scientific Memberships	American Public Health Association
Journal Referee	**Emerging Infectious Diseases, Parasites & Vectors, American Medical Informatics Association (AMIA)**
Interests	Infectious Diseases, Public Health Informatics, Bioinformatics

Software

April 2018	kdtApp: A small Shiny app for the estimation of Knock-Down Time based on aggregated WHO or CDC bottle Bioassay data. URL: https://razondekon.shinyapps.io/kdtApp/
Feb 2018	Authorvis: a co-authorship network exploration, and link prediction tool specific to the scientific collaborative research network of Malaria, Tuberculosis and HIV/AIDS in the Republic of Benin. URL: https://hub.docker.com/r/rosericazondekon/authorvis, https://github.com/rosericazondekon/authorvis

Github repository: https://github.com/rosericazondekon

List of Publications

1. **Roseric Azondekon**, Zachary James Harper, Fiacre Rodrigue Agossa, Charles Michael Welzig, and Susan McRoy. *Scientific Authorship and Collaboration Network Analysis on Malaria Research in Benin: Papers Indexed in the Web of Science (1996–2016)*. Global Health Research and Policy 3, no. 1 (April 6, 2018). doi:10.1186/s41256-018-0067-x.

2. Fiacre R Agossa, Virgile Gnanguenon, Rodrigue Anagonou, **Roseric Azondekon**, Nazaire Aïzoun, Arthur Sovi, Frédéric Oké-Agbo, Michel Sèzonlin, Martin C Akogbéto: *Impact of Insecticide Resistance on the Effectiveness of Pyrethroid-Based Malaria Vectors Control Tools in Benin: Decreased Toxicity and Repellent Effect*. PLoS ONE 12/2015; 10(12):e0145207., DOI:10.1371/journal.pone.0145207
3. Rodrigue Anagonou, Fiacre Agossa, **Roseric Azondékon**, Marc Agbogan, Fréderic Oké-Agbo, Virgile Gnanguenon, Kèfilath Badirou, Ramziath Agbanrin-Youssouf, Roseline Attolou, Gil Germain Padonou: *Application of Polovodova's Method for the Determination of Physiological Age and Relationship between the Level of Parity and Infectivity of Plasmodium falciparum in Anopheles gambiae ss, South-eastern Benin*. Parasites & Vectors 12/2015; 8(1-1):117., DOI:10.1186/s13071-015-0731-7
4. Martin C Akogbéto, Rock Y Aïkpon, **Roseric Azondékon**, Gil G Padonou, Razaki A Ossè, Fiacre R Agossa, Raymond Beach, Michel Sèzonlin: *Six years of experience in entomological surveillance of indoor residual spraying against malaria transmission in Benin: Lessons learned, challenges and outlooks*. Malaria Journal 06/2015; 14(1):242., DOI:10.1186/s12936-015-0757-5
5. Virgile Gnanguenon, Fiacre R Agossa, Kefilath Badirou, Renaud Govoetchan, Rodrigue Anagonou, Fredéric Oke-Agbo, **Roseric Azondekon**, Ramziath AgbanrinYoussouf, Roseline Attolou, Filemon T Tokponnon, Rock Aïkpon, Razaki Ossè, Martin C Akogbeto: *Malaria vectors resistance to insecticides in Benin: Current trends and mechanisms involved*. Parasites & Vectors 04/2015; 8(1):223., DOI:10.1186/s13071-015-0833-2
6. Rodrigue Anagonou, Fiacre Agossa, **Roseric Azondékon**, Marc Agbogan, Fréderic Oké-Agbo, Virgile Gnanguenon, Kèfilath Badirou, Ramziath Agbanrin-Youssouf, Roseline Attolou, Gil Germain Padonou, Arthur Sovi, Razaki Ossè, Martin Akogbéto: *Application of Polovodova's method for the determination of physiological age and relationship between the level of parity and infectivity of Plasmodium falciparum in Anopheles gambiae s.s, south-eastern Benin*. Parasites & Vectors 01/2015; 8:117.
7. Rodrigue Anagonou, Fiacre Agossa, Virgile Gnanguenon, Bruno Akinro, Gil Germain Padonou, Renaud Govoetchan, Rock Aikpon, Arthur Sovi, **Roseric Azondekon**: *Development of new combined method based on reading of ovarian tracheoles and the observation of follicular dilatations for determining the physiological age of Anopheles gambiae ss*. 01/2015; 9(1):9-15.
8. Virgile Gnanguenon, Renaud Govoetchan, Fiacre R Agossa, Razaki Ossè, Frédéric Oke-Agbo, **Roseric Azondekon**, Arthur Sovi, Roseline Attolou, Kefilath Badirou, Filémon T Tokponnon, Gil G Padonou, Martin C Akogbeto: *Transmission patterns of Plasmodium falciparum by Anopheles gambiae in Benin*. Malaria Journal 11/2014; 13(444)., DOI:10.1186/1475-2875-13-444
9. Nazaire Aïzoun, **Roseric Azondekon**, Rock Aïkpon, Virgile Gnanguenon, Razaki Osse, Alex Asidi, Martin Akogbéto, Asian Pac, J Trop Biomed: *Study of the efficacy of a Wheaton coated bottle with permethrin and deltamethrin in laboratory conditions and a WHO impregnated paper with bendiocarb in field conditions Asian Pacific Journal of Tropical Biomedicine*. Asian Pacific Journal of Tropical Biomedicine 06/2014; 4(6):492-497., DOI:10.12980/APJTB.4.2014C1111
10. Nazaire Aïzoun, Virgile Gnanguenon, **Roseric Azondekon**, Rodrigue Anagonou, Rock Aïkpon, Martin Akogbéto: *Status of organophosphate and carbamate resistance in Anopheles gambiae sensu lato from the Sudano Guinean area in the central part of Benin, West Africa*. 04/2014; 8(4):61-68., DOI:10.5897/JCAB2013.0400

11. Nazaire Aïzoun, Rock Aïkpon, **Roseric Azondekon**, Alex Asidi, Martin Akogbeto: *Comparative susceptibility to permethrin of two Anopheles gambiae s.l. populations from Southern Benin, regarding mosquito sex, physiological status, and mosquito age*. 04/2014; 4(4)., DOI:10.12980/APJTB.4.2014C1093
12. Renaud Govoetchan, Virgile Gnanguènon, Euloge Ogouwalé, Frédéric Oké-Agbo, **Roseric Azondékon**, Arthur Sovi, Roseline Attolou, Kefilath Badirou, Ramziyath Agbanrin Youssouf, Razaki Ossè, Martin Akogbéto: *Dry season refugia for anopheline larvae and mapping of the seasonal distribution in mosquito larval habitats in Kandi, northeastern Benin*. Parasites & Vectors 03/2014; 7(1):137., DOI:10.1186/1756-3305-7-137
13. Nazaire Aïzoun, Rock Aïkpon, **Roseric Azondekon**, Virgile Gnanguenon, Razaki Osse, Gil Germain Padonou, Martin Akogbéto: *Centre for Disease Control and Prevention (CDC) bottle bioassay: A real complementary method to World Health Organization (WHO) susceptibility test for the determination of insecticide susceptibility in malaria vectors*. DOI:10.5897/JPVB2013.0144
14. Nazaire Aïzoun, **Roseric Azondekon**, Rock Aïkpon, Rodrigue Anagonou, Virgile Gnanguenon, Martin Akogbéto: *Dynamics of insecticide resistance and exploring biochemical mechanisms involved in pyrethroids and dichlorodiphenyltrichloroethane (DDT) cross-resistance in Anopheles gambiae s.l populations from Benin, West Africa*. 03/2014; 8(3):pp. 41-50., DOI:10.5897/JCAB2014.0406
15. Arthur Sovi, Innocent Djègbè, Lawal Soumanou, Filémon Tokponnon, Virgile Gnanguenon, **Roseric Azondékon**, Frédéric Oké-Agbo, Mariam Okè, Alioun Adéchoubou, Achille Massougbodji, Vincent Corbel, Martin Akogbéto: *Microdistribution of the resistance of malaria vectors to Deltamethrin in the region of Plateau (southeastern Benin) in preparation for an assessment of the impact of resistance on the effectiveness of Long Lasting Insecticidal Nets (LLINs)*. BMC Infectious Diseases 02/2014; 14(1):103., DOI:10.1186/1471-2334-14-103
16. Renaud Govoetchan, Virgile Gnanguenon, **Roseric Azondékon**, Rodrigue Fiacre Agossa, Arthur Sovi, Frédéric Oké-Agbo, Razaki Ossè, Martin Akogbéto: *Evidence for perennial malaria in rural and urban areas under the Sudanian climate of Kandi, Northeastern Benin*. Parasites & Vectors 02/2014; 7(1):79., DOI:10.1186/1756-3305-7-79
17. Virgile Gnanguenon, **Roseric Azondekon**, Frederic Oke-Agbo, Raymond Beach, Martin Akogbeto: *Durability assessment results suggest a serviceable life of two, rather than three, years for the current long-lasting insecticidal (mosquito) net (LLIN) intervention in Benin*. BMC Infectious Diseases 02/2014; 14(1):69., DOI:10.1186/1471-2334-14-69
18. Fiacre R Agossa, Rock Aïkpon, **Roseric Azondékon**, Renaud Govoetchan, Gil Germais Padonnou, Olivier Oussou, Frédéric Oké-Agbo, Martin C Akogbéto: *Efficacy of various insecticides recommended for indoor residual spraying: Pirimiphos methyl, potential alternative to bendiocarb for pyrethroid resistance management in Benin, West Africa*. Transactions of the Royal Society of Tropical Medicine and Hygiene 02/2014; 108(2):84-91., DOI:10.1093/trstmh/trt117
19. **Roseric Azondekon**, Virgile Gnanguenon, Frederic Oke-Agbo, Speraud Houevoessa, Michael Green, Martin Akogbeto: *A tracking tool for long-lasting insecticidal (mosquito) net intervention following a 2011 national distribution in Benin*. Parasites & Vectors 01/2014; 7(1):6., DOI:10.1186/1756-3305-7-6
20. Nazaire Aïzoun, Rock Aïkpon, **Roseric Azondekon**, Alex Asidi, Martin Akogbéto, Asian Pac, J Trop Biomed: *Comparative susceptibility to permethrin of two Anopheles gambiae s.l. populations from Southern Benin, regarding mosquito sex, physiological status, and mosquito age Asian Pacific Journal of Tropical Biomedicine*.

21. Nazaire Aïzoun, Rock Aïkpon, Virgile Gnanguenon, **Roseric Azondekon**, Frédéric Oké-Agbo, Gil Germain Padonou, Martin Akogbéto: *Dynamics of insecticide resistance and effect of synergists piperonyl butoxide (PBO), S.S.S-tributylphosphorotrithioate (DEF) and ethacrynic acid (ETAA or EA) on permethrin, deltamethrin and dichlorodiphenyltrichloroethane (DDT) resistance in two Anopheles gambiae s. l. populations from Southern Benin, West Africa*. DOI:10.5897/JPVB2013.0137
22. Renaud Govoetchan, Arthur Sovi, Rock Aïkpon, **Roseric Azondékon**, Abel Kokou Agbévo, Frédéric Oké-Agbo, Alex Asidi, Martin Akogbéto: *The effects of oviposition site deprivation up to 40 days on reproductive performance, eggs development, and ovipositional behaviour in Anopheles gambiae (Diptera, Nematocera, Culicidae)*. DOI:10.5897/JPVB2013.0136
23. Arthur Sovi, **Roseric Azondékon**, Rock Y Aïkpon, Renaud Govoétchan, Filémon Tokponnon, Fiacre Agossa, Albert S Salako, Frédéric Oké-Agbo, Bruno Aholoukpè, Mariam Okè, Dina Gbénou, Achille Massougbodji, Martin Akogbéto: *Impact of operational effectiveness of long-lasting insecticidal nets (LLINs) on malaria transmission in pyrethroid-resistant areas*. Parasites & Vectors 11/2013; 6(1):319., DOI:10.1186/1756-3305-6-319
24. Virgile Gnanguenon, **Roseric Azondekon**, Frederic Oke-Agbo, Arthur Sovi, Razaki Ossè, Gil Padonou, Rock Aïkpon, Martin C Akogbeto: *Evidence of man-vector contact in torn long-lasting insecticide-treated nets*. BMC Public Health 08/2013; 13(1):751., DOI:10.1186/1471-2458-13-751
25. Nazaire Aïzoun, Razaki Ossè, **Roseric Azondekon**, Roland Alia, Olivier Oussou, Virgile Gnanguenon, Rock Aikpon, Gil Germain Padonou, Martin Akogbéto: *Comparison of the standard WHO susceptibility tests and the CDC bottle bioassay for the determination of insecticide susceptibility in malaria vectors and their correlation with biochemical and molecular biology assays in Benin, West Africa*. Parasites & Vectors 05/2013; 6(1):147., DOI:10.1186/1756-3305-6-147
26. Gil Germain Padonou, Ghelus Gbedjissi, Anges Yadouleton, **Roseric Azondekon**, Ossé Razack, Olivier Oussou, Virgile Gnanguenon, Aikpon Rock, Michel Sezonlin, Martin Akogbeto: *Decreased proportions of indoor feeding and endophily in Anopheles gambiae s.l. populations following the indoor residual spraying and insecticide-treated net interventions in Benin (West Africa)*. Parasites & Vectors 11/2012; 5(1):262., DOI:10.1186/1756-3305-5-262
27. HM Gbetoh, PA Edorh, MF Gouissi, AS Hounkpatin, **R Azondekon**, P Guedenon, L Koumonlou, FS Loko, M Boko: *Dosage du plomb et du cadmium dans le sperme des sujets consultant pour infertilité masculine dans la ville de Cotonou*. International Journal of Biological and Chemical Sciences 08/2012; 6(2-2):582., DOI:10.4314/ijbcs.v6i2.3
28. Padonou G G, Sezonlin M, Gbedjissi G L, Ayi I, **Azondekon R**, Djenontin A, Bio-Bangana S, Oussou O, Yadouleton A, Boakye D, Akogbeto M: *Biology of Anopheles gambiae and insecticide resistance: Entomological study for a large scale of indoor residual spraying in south east Benin*. DOI:10.5897/JPVB11.018

Conference Proceedings

1. **Roseric Azondekon**, Zachary Harper, Charles Welzig: *Scientific Authorship and Collaboration Network Analysis on Malaria Research in Benin: papers indexed in the Web Of Science (1996-2016)*. American Society of Tropical Medicine and Hygiene 66th Annual Meeting, Baltimore, MD, USA; 11/2017

2. Zachary Harper, **Roseric Azondekon**, Charles Welzig: *Exploring Connectivity Dynamics using Deep Neural Network Models from Magnetoencephalographic Data*. Organization for Human Brain Mapping, Vancouver, BC, Canada; 06/2017
3. Samuel A. Abariga, **Roseric Azondekon**, Krishna C. Poudel: *Epidemiology of Hepatitis C and HIV co-infection in Asia*. 143rd American Public Health Association Annual conference, Chicargo IL; 11/2015
4. **Roseric Azondekon**, Samuel Abariga, Krishna C. Poudel: *Prevalence of Human Immunodeficiency Virus and Hepatitis C virus co-infection in Africa: A Systematic Review and Meta-analysis*. 142nd APHA Annual Meeting & Exposition, New Orleans, LA, USA; 11/2014
5. Akogbeto C. Martin, Osse Razaki, **Azondekon Roseric**, Yadouleton Anges: *Implementing full coverage of long lasting insecticidal nets: a good alternative strategy after cessation or abandon of indoor residual spraying*. ASTMH 61st Annual Meeting, Atlanta, Georgia, USA; 11/2012
6. **Roseric Azondekon**: *Entomological profile and cartography of malaria vector insecticide resistance in Benin*. 5th MIM Pan-African Malaria Conference, Narobi, Kenya; 11/2009