NETWORK ANALYSIS OF RICE HEALTH SUREVEY DATA FOR CHARACTERIZATION OF YIELD REDUCING FACTORS AND YIELD LIMITING FACTORS OF TROPICAL RICE ECOSYSTEM IN SOUTH AND SOUTHEAST ASIA

SITH JAISONG

THESIS OUTLINE SUBMITTED TO THE FACULTY OF GRADUATE SCHOOL
UNIVERSITY OF THE PHILIPPINES LOS BAÑOS
IN FULFILLMENT OF THE
REQUIREMENT FOR THE
DEGREE OF

PH.D OF SCIENCE
(Plant Pathology)

May 2015

TABLE OF CONTENTS

List of Figures

CHAPTER I

INTRODUCTION

The threats of pests and diseases to global rice production are significant yield reducing factors. Additionally, future rice production will need to grow by 2.4% per year in order to meet the demands of a growing population (?, ?). Addressing these yield-reducing factors is essential for food security in rice consuming societies now and in the future.

Description of pest management is complicated because human activities regarding with agricultural practices to the pests are diverse. To achieve this, the knowledge of relationships between pests and human activities including environmental factors are needed to untie (?, ?). While these studies are important in characterization the of pattern of pest injuries and the pattern of components relating to production situation, and how it related to each others. They have not been insufficient for explanation about change of relationships of injuries and components of production situation in term of time and regions, or how such relationships may change with climate variations. For instance, the relations between insect pest and weed were found only in the certain time, or certain locations. Over the long run, in season correlation between the frequency of patterns of pest incidence at different locations are also important to design the pest management strategies.

We live in an increasingly connected era where by early 2013 over 90% of all data had been generated to that point (?, ?). In agriculture we should strive to find ways to harness our own ”big data”. We have a daunting task ahead of us that will require a myriad of approaches from well-know and understood to new approaches like big data analytics. Using new approaches to analyze so called “big data”, we can start making new discoveries in relationships between factors of which we were previously unaware. These newly discovered relationships could be useful in designing and developing methods for managing plant pests and diseases.

Network models provide powerful tools in many branches of science, which are applied for generate to . A network is an abstract model composed of a set of nodes or vertices, a set of edges, links or ties that connect the nodes, together with information concerning the nature of the nodes and edges. The nodes usually represent entities and the edges represent their relations. This simple model can be used to describe many kinds of phenomena, such as social relations, technological and biological structures, and information networks.

OBJECTIVES

My objectives are to synthesize approaches to characterize the yield constrains under agroecosystem. I also address how pests and cropping practices relate to yields individually and simultaneously. For better understanding, climatic variable can be possibly added. Finally, I propose the examples of new applications and a conceptual framework for exploring the relationships of injuries caused by pests, human activities under different geographic location.

The networks constructed will attempt to answer the following question.

1. How can the rice yield losses be examine from the perspective of networks analysis. What the key factors affect to rice yield productivities? and how variation is in different locations?

2. What relationships between components of networks as defined in the variables in survey profile data?

3. How are theses relationships affects by different locations ?

The anticipation of the proposed work will provide the insights of rice injuries from network inference of rice crop health survey data. Complete an analysis of the crop health survey data using the new network model and provide visualizations and interpretations of the results and make rice crop health recommendations based on findings. It will be very helpful for plant health authorities worldwide, in order to design specific strategies for rice pest and disease management and to limit the impacts of these yield reducing factors.

CHAPTER II

REVIEW OF LITERATURE

Introduction

The applications of network analysis have increased exponentially over the past two decades in various disciplines. Even though documented applications of network analysis in plant pathology are still relatively sparse, network applications in the social science, systems biology and ecology have been increasingly found. ? (?, ?, ?, ?) presented useful concepts and methods of network analysis in the studies related to plant pathology. I review the empirical works that exist and argue that network analysis is a promising approach for exploring the questions in the context of plant pathology.

This chapter contains four sections of review of network analysis and its applications. In the first sections, I introduces a brief overview of the concepts and methods of network analysis. I then discuss the new dimensions of network analysis that are not found in other approaches. In the first two sections, I provide an overview of key of network concepts and place network analysis as a methodology within the broader toolbox of methods commonly used in biological studies.

In the third section I scope network analysis into the current applications of plant pathological research, particularly in plant disease epidemiology and molecular plant pathology, in which network analysis has been broadly applied, and increasingly documented. It provides fruitful tools for visualizing, analyzing and understanding complex relationships in the studies of plant disease. For instance, network models of genes or proteins pertaining plant defense mechanism and network models revealing spatial distribution of plant disease through trade networks were reviewed by ? (?) and ? (?) respectively.

The last section presents analytical techniques and strategies to apply network analysis for studying pest management. It introduces three strategies of network applications, which are developed and applied in the studies of systems biology and ecology, and concludes the brief discussions of potential applications to the studies of crop heath management that have yet been undertaken.

PartI: Network Analysis

Introduction to network analysis

Network analysis is applied for determining relationships between elements of interest. It offers toolkits for visualizing data in a network model and measuring its properties, and network thinkings (?, ?). It been widely used by various branches of science, such as social science, ecology, biology, computer science, and many others to study the interactions between elements, e.g., the relationships of students in school (?, ?), species in food webs (?, ?), interactions of genes or proteins in cells (?, ?), or the connections of computer in the network (?, ?, ?).

? (?) loosely categorized four types of networks based on different complex data. The first category is social network, representing sets or groups of people forming some patterns of contacts or interactions between them such as the patterns of friendship or business relationships. Analyzing the structure of whole social entities gives us the perspectives from a social network, which enables us to explain the patterns observed. ? (?) analyzed the social behaviors in high school students using social network approaches. (?, ?) applied network analysis to compare the social relationships and friendships between children with and without Autism spectrum disorder (ADS). The second type of network is an information network or knowledge network. The classic example of this network is the network of citations between academic papers (?, ?). The articles cited other papers, which have related topics. They formed a citation network that has vertices as articles and direct links as citations. The citation network visualizes the structure and the movement of the information. The third category, technological network, is object connected network, or man–made network which represents a physical connection between objects. This network is mostly applied for illustrating physical structures and systems such as the electrical power grid, the connections of rivers, transport systems, etc. The fourth category of network is a biological network. It represents the biological systems such as genes to genes, genes to protein, protein to protein interactions, which enable biologists understand the connections and interactions between individual constituents including genes, proteins, and metabolites at the level of the cell, tissue and organ to ultimately describe the entire organism system. Biologists use biological networks in various branches of biology at different levels (from a single molecule to an entire organism). For example, (?, ?, ?) studied in the patterns of gene expression in different conditions and different types of cells (normal cells and cancer cells) in order to characterize the genes that change and do not change following the particular conditions; (?, ?) applied a molecular ecological network analysis to study the communities of soil microorganisms. Networks revealed the complex relationships between microbial species in soils and their communities. Moreover, network analysis enables ecologists to understand ecological properties and predict the ecological roles of species in a soil ecosystem. Although the application of each type of network approach varies, all four categories of networks share a common empirical focus on relational structure and a similar set of mathematical analysis.

In plant pathological studies network analysis can be the powerful tool to study plant disease epidemics. ? (?) showed the the potential for the use of network analysis in plant epidemiology. Network analysis can reveal the dynamics of the disease spread. The theory and tools of network analysis support plant pathological studies. ? (?) reviewed the studies related to plant disease which applied network analysis. Networks applied in the studies of plant disease are generally similar to ecological studies. In plant pathology, most network analysis is used in molecular plant pathology as a model of the interactions between genes and/or proteins, and other cellular constituents contributing to host plant resistance or pathogen infection. Spatial analysis

Concepts, principles, and methods of network analysis

A network represents relationships between of elements of interests, which is defined by links (edges) among nodes (vertices). Nodes can be units of interests or studies, and links represent interactions between nodes. Network analysis aims the association among nodes rather than the attributes of particular nodes. In network analysis, networks are defined as any set or set of links between any set or sets of nodes.

Network analysis follows three principles. Nodes and their behaviors are mutually dependent, not autonomous; links between nodes can be channels for transmission of both material (for example, money, disease) and non-material (for example, information, knowledge, relationship, interaction) and; persistent pattern of association among nodes create structure that can define, enable, or restrict the behavior of a node.

Network models have two different organizational structures depending on the goal of the representation and analysis (?, ?). Flow models view the network as a system of pathways along which things flow between nodes. Analysis of flow models can, for example, identify which nodes in the network are more active, or which ones are more powerful. Flow models are good for evaluating processes, as was shown in these reviews of plant disease spread (?, ?, ?). Architectural models tend to focus on the structure of the network, seeking to discern whether specific structures lead to similar outcomes, or whether actors in similar network positions behave in similar ways. Ecological applications related to the ecology and spatial structure of ”community” tend to be organized and analyzed as architectural models. For example, ? (?) studied the networks of soil microbial interactions. Network models describes how microbial populations change over time, which will require the use of dynamic models of microbial communities. Beyond these basic principles, network analysis enables the calculation of structural properties of nodes, groups, or the entire network.

Measuring network properties

A network is made up of nodes and links from relational data. It is constructed from adjacency matrix, which is obtained from analysis using metric algebra techniques. The row and column headings for an adjacency matrix are identical, listing the names of the components involved in the network. In the simplest case, the cells of the matrix are coded with ”1” if an link exists between the node or ”0” if no edge exists. However, a link can be valued. Value indicates a characteristic of the relationship that the research has quantified. The values may be binary, such as whether two friends recognize each other, or variable strength, e.g., the number of mutual friends between two friends. Network link need not to imply positive or cooperative interaction; they can also be negative or competitive interaction between two individuals.

The distribution of links in a network suggests two important structural characteristics: centrality (importance) of nodes in the network and division of the network into subgroups. Variants of centrality in a network include degree, closeness, and betweenness. Degree centrality of a node is the sum of the value of the links between that node and every other node in the network. This measure tells us how well-connected a particular node is to the other nodes. Closeness centrality is calculated using the length of the path between a node and every other node. This measure could estimate the time required for information or resources to propagate to a given node in a network. Betweenness centrality corresponds to the number of paths in the network that pass through a particular node, and therefore measures the dependence of a network on a particular node for maintaining connectedness (?, ?). ? (?, ?, ?) are recommended references for descriptions of the theory and uses, as well as the formal calculation of these measures.

Part II: The dimensions of network analysis unlike other approaches

With an increasing trend towards a systems level perspective in the science, away from the reductionism that characterized much of the previous century (?, ?), the development of network analysis challenges conventional approaches to uncover the complex patterns of interaction or relationships.

The difference of network analysis from conventional approaches is that it involves different methods of analysis. Network analysis models the relational data, and measures various properties of network. A challenge of network analysis is the consideration of the properties of interactions between elements in the system, which interactions are assumed that they are dependent on each other. That is, when element A has a relationship with element B, the relationship is not only considered to be independent of element A and B, but also relationship of element A and B to be independent to other element. Unlike network analysis, traditional research methods consider attributes such as variables in a wide variety of statistical analyses, these methods are sometimes referred to as variable analysis (?, ?).

As objects for representing interactions among elements of a complex system, network graphs are primarily focused in network analysis. Interactions or relationships are represented as edges in networks. They present only when interactions are exist between nodes. Moreover, network concepts place nodes in rational distance according to the level of relationships. For example, to visualize the strong relationship of A and B, network places the node represented A near the node represented B. Here is the new dimensions network of visualizing interactions or relationships.

Extending from considering the interaction of two nodes, network graphs illustrate nicely the connections. When the element A associated with the element B, and also with the element, the network graph is able to visualize the connection between node A to node B, and to node C. This pattern, so-called ”betweenness”, potentially is useful to represent what the important node is in network graph because node A has high connectivity and is a connector between node B and node C. For example, In protein-protein interaction (PPI) network, proteins with high betweenness have been termed ”bottlenecks”, for their role as key connector proteins with essential functional and dynamic properties (?, ?).

The heart of network study is structure; network processes replace vertices and edges in rational spaces depending on the selected graph layout methods. The network structure enable scientists to study interactions with behaviors and attributes of pairs of vertices. Networks would reveal the groups of vertices, which are closely related or similar to each other. This is also similar to the basis for the cluster analysis. For example, if the employees of the same company may share similar attributes such as location or educational background and they are close in network. When the relationships are simple and the differences in node attributes are clear, the conventional analytic approaches such as cluster analysis, principal component analysis, correspondence analysis are sufficient. However, when relationships are complex or vertices attributes are more nuanced, clear answers using conventional analysis may prove elusive. Network analysis offers a tool to help researchers disentangle some of the relational complexities.

Part III: Network and botanical epidemiology

Recently, a broad expansion of applications of network analysis has occurred across many disciplines over the past decade, and several researchers have evaluated the impact and potential of network approach on their disciplines. Plant pathology embrace network analysis to dominantly study in plant disease spreads and plant-pathogen interactions. ? (?, ?) provide insightful analysis of the reasons why network analysis is useful in the study of plant pathology. With its generality and flexibility, network representation can be used at a variety of levels in plant pathology, from biological mechanisms of plant–pathogen interactions, to the development of plant disease spreads among fields, farms, and landscapes and to trade movement of plants infected by pathogens or infested by insects among regions and countries.

There are two fields in the literature relevant to plant pathology, which present particularly strong growth and prove that network analysis has significant potential to augment traditional analysis methods. The first is plant disease epidemiology, which investigates questions related to plant disease spread. The second is plant molecular biology, which investigate question related to biological network. Here I briefly review these two themes and identify a few studies that have begun to integrate them.

Using Network analysis to understand plant disease spread

The idea of plant disease spread is that the probability of infection embedded in the connection or the contact patterns between susceptible/infected plants, and it forms as the networks. For network graphs, plant disease spread and establishment are able to be modeled in directed or indirected networks using scenarios concerning networks (?, ?). The disease starts at a single node, and will connect to other nodes when they are infected over time with a certain probability of transmission. In turn, an infected nodes will infect at the next time step depending on their infection status and on a certain probability of persistence. The probability of infection transmission is the same for all connections in a given network replicate. The probabilities of persistence and transmission define an epidemic threshold, which is independent of the starting node of the epidemic, and will influence to the structure of network.

Network analysis includes models developed specifically to answer questions related to flows of information and structures of connections through networks. Tracing information flows through a trade network can expose critical gaps or inefficiencies that may contribute to plant disease control policies as classified by. Routes and network nodes of transmission were studied for trade transmission. ? (?) revealed *Phytophthora ramorum* epidemics in the horticultural trade network. For P. *ramorum*, the epidemic networks in the horticultural trade and the semi-natural landscape showed independent systems. Combining genetic network analysis and data on trace forward and trace back on movement of plants nursery trade supported to identified confidentially P. *ramorum* migration. From this approaches, it was clear that pathogen was introduced originally from nurseries, which P. *ramorum* populations in nurseries are genetically ancestral to all Californian forest populations.

Network models developed enable the researcher, policy makers or related person to identify or predict what entities play the key role in the network. In trade network. network analysis suggests broadly that it would be sensible to place quarantine efforts on hubs or on connections between major hubs. However, wherever effort is placed, a disproportionate increase in quarantine effort is needed to keep the rate of flow of pathogens across trade links constant as the trade through links increases (?, ?, ?, ?). For instance, trade network of plants and plant products across the world and within countries give the picture on how to be able to control the flow of pathogens. The strategy should be designed by focusing on links to and from hubs, nodes which have high degree of connectivity would increase efficiency to achieve control plant pathogen spreads. To cope with increasing volumes in trade of potential infected plants, this insight may be very helpful for plant heath authorities target at the traders who have high connection activities or find the major pathways. The control of disease or quarantine can be made more efficient and effective.

Using Network analysis to understand molecular plant pathology

The development of bioinformatics and biostatistics method and increasing number of biological data have influence to the ”new” biology (?, ?). Recently, the representation of biological systems by networks (graphs) is commonly applied in biology to analyze the systemic interplay of biological components. When the systems (biological systems, for example) are very large and have many fractions of unknown contents in there, top-down approaches are often applied (?, ?). To apply network analysis attempts to infer properties of the system (such as network structure or parameters that can capture the system dynamics range interactions over time) from large scale biological data sets in order to identify interesting features that may be tested using more targeted experiments.

Network analysis offers tools to visual the myriad information. ? (?) built a network from the gene-for-gene relationships between rice and various avirulence genes of the pathogen *Xan-*

thomonas oryzae, which its nodes represented isogenic lines of rice and links are connected if they share genes with high resistance (with respect to avirulence genes). This network can help rice breeder to identifying particularly promising genes for developing host resistance to pathogens. General presentation, such as genetic maps, frequency distributions, etc. are also equally value, but network models are clearly able to provide an overview of relationships in a given system.

The network approach focuses on components of network structure that cannot be created from observing individual nodes alone. Recently, biologists have attempted to understand the nature and consequence if biological complexity using network analysis to understand the network structure of biological. For instance, networks can be constructed from available data for a certain plant pathogen from multiple locations/hosts based on the similarity among the pathogen strains. ? (?) showed the good examples, which are co-occurrence network of the *Phytophthora ramorum* infected plant genera different environment.The networks may be helpful in identifying host taxa playing a important role in spreading a certain disease in the semi-natural environment, in crop plants, and plants in the trade.

Network analysis can contribute to our understanding of biological mechanism and interaction, and influence by analyzing the network composition and interdependent relationship (network structure) of elements in biological processes. Granted, network analysis is not needed for simple assessment of network composition; that is, to measure the relative levels of key components in the processes. However, what it does offer is the ability to identify and compare the structural positions of individuals and their relationships. Systematically and simultaneously analyzing network composition and structure provides much deeper insights in holistic view. Plant-pathogen interaction network by ? (?) revealed the complicated that a large number of novel Arabidopsis protein-pathogen effector interactions, provided evidence that pathogen effectors target a limited number of host immune proteins, and demonstrated that effectors from very distantly related pathogens interact with the same host proteins. (?, ?) construct the network of soil fungal community. They compare the fungal networks with different condition, yield-invigorating and yield-debilitating soils under prolonged potato monoculture. from this comparative network, the result showed that *Sordariales* and *Hypocreales* were major affected phylogenetic group. Network analysis enables the identification of elements and their relationships that need to be bridged to overcome problems of relational complexity.

Part IV: Application of Network analysis to study of Agro-Ecosystems

The usefulness of network analysis related to applications of plant disease management. I mentioned in the previous topic falls into two broad types of applications. The first is network application in landscape plant disease epidemiology . This applications tend to be more flow oriented, while the second type of networks related to plant molecular biology tends to be more emphasized in the structure. In both of these applications the literature shows a progression tend to from theoretical to empirical approaches over time (?, ?). For this chapter I give the literature related to specific applications as the third type of plant disease studies, which network analysis is used for and focus on analyzing plant pathosystem.

The network theory applied in plant pathology would broaden windows of opportunities to plant pathologists to understand the causality of plant disease epidemics. Diversity is primary causes, for example to shape plant disease epidemics. (?, ?) claimed that networks is a tool to examine the relationships between plant diversity and ecosystem susceptibility to plant pathogens, and networks also show the capability of revealing system perturbations, which we can see what the difference between system with and without interference(e.g. plant pathogen introduction or pesticide use), so they will enable us to predict the consequences in order to improve the sustainability of agriculture.

Alternatively, network analysis can be implemented to analyze plant pathosystem. To maintain the agricultural sustainability, plant pathologists need to develop a more holistic approach to crop protection. It is pathosystem analysis, which considers the various pests, including disease, insects and weed that affect a crop, and cropping practices, and also determines their interactions. A survey may provide the necessary overview of the pathosystem; adequate methods for analyzing survey data can produce preliminary information on its behavior including major interactions.In this context, surveys can be considered as part of a systems approach. use the survey data to analyze the characterization, the dynamics or the behavior of pathosystem.

Traditional approaches are found limitation in analysis the data with high heterogenouse and dimensionality. The survey data have that format. Various statistical approaches are used to reduce dimensions and extract major feature, including principal components, correspondence analysis The survey generated both of quantitate and qualitative data. In multivariate statistical analysis, traditional methods analyzing the data set which contained both type of variable are principle component analysis, correspondence analysis etc. Correspondence analysis normally have been picked up to analyze the survey data related to plant disease (?, ?, ?, ?). Briefly, there are two main stages. The first stage is to convert the quantitative variables into classes by using cluster analysis, then perform Chi-square test to prove that the class of variables are independence. The number of classes and their limits could be fixed in line with critical thresholds known beforehand (?, ?). The second stage is to perform correspondence analysis. The limits of the categories of explanatory variables used in the segmentation played the role of decision thresholds. because this methods has a limit, which there is no solid role of choosing category limits, ? (?) claimed that if categorized variables have influence experimentally or particularly to a large degree, those limits had been chosen. The logic of the results obtained tended to back the choice made. Although no general rule can be imposed when choosing category limits (?, ?), it is necessary to validate them experimentally, particularly if they have to influence decision making, such as the decision to carry out chemical control.

Like multivariate statistical analysis, network analysis are equipped with a set of tools to analysis the large datasets with homogenous and heterogenous. In biological point of view, correlation-based networks are commonly used for generating biologically meaningful hypotheses and for gaining novel insights into the complex relationship (?, ?). Expending the key role of the classic disease triangle (host–pathogen–environment), plant pathosystem profoundly conserved that classic role. Moreover, human and human activities have possibly additional factor, and strongly influence plant disease epidemics. For plant disease control (may included pests) to succeed, the holistic view of the interconnectivity of components causing plant disease development and their influences is critical. This section reviews the challenges of network application for analyzing plant pathosystem that have not been discussed previously in this literature. These challenges discussed fall into three concept of network analysis.

Concept 1: Networks are generated from multiple pairwise relationships of data through the statistical approaches. The networks are inferred from the assumption of interaction and influence between components generated from statistical relation in the observed data. Interaction between two elements involved falls into one type of interaction in someway. Otherwise, there will not have interaction. In microbial association network, the relationships can be predicted from determination to co-occur or show a similar pattern over the multiple samples or time. Also similarly in context of protein-protein interaction, for example, if protein A induces expression of protein B, then expectedly levels of protein B are high whenever levels or specific molecular states of its activator A are high. The reverse of this logic is that statistical correlation between protein states indicates a potential interaction between them.

A relationship between two entity of network can be predicted by means of two group of network inference methods; Pairwise relationships: similarity-base network inference and Complex relationships: regression- and rule-based networks.

There are serval techniques to determine the similarity between variables depending on the biological question to be answered. Mainly using either Pearson's or Spearman's correlations for data (?, ?, ?) and the hypergeometric distribution for presence-absence data (?, ?). Another popular similarity-based network inference methodology is local similarity analysis (LSA) (?, ?), which can detect similarity between shifted abundance profiles and is therefore frequently used to build association networks from time series data. In an interesting alternative approach, Pearson correlation thresholds are determined using random matrix theory (?, ?). These approaches should be evaluated, and compared which approaches are good evaluation. Pairwise relationship does not always mean that there is a biological relations for two compared variables because they may be influenced by other variables in the system. Regression- and rule base networks are also applied, especially in ecology fields to infer complex interactions (?, ?), and there are other techniques such as mutual information analysis must be applied (?, ?).

Concept 2: Networks are not static, but dynamic according to the given data and environments. Networks can dynamically respond and adapt to the internal state and external signals (?, ?). As the internal state, backgrounds of nodes have big influence to the structure and behavior network, and give rise to significant difference across individuals. Backgrounds are included information about where the data are from. The different sources and different times collecting the data strongly determine the network behaviors.

Exogenous signals such as nutrients, chemicals, environmental conditions, affect to networks. Networks can be constructed from different sets of data (e.g., from different locations, sources), structural properties can be used to determine the differences between the networks (?, ?). Protein-protein interaction network models can vary significantly between different condition (?, ?). Comparative network broadly are construct for comparing the series of parameters related to network topologies under different situation. For examples, in a soil microorganism community, the context can profound impact on how the communities at difference between different types of soil. (?, ?) compared the fungal communities of yield-invigorating and -debilitating soils from prolonged potato monoculture. Fungal network of healthy soil showed high connectivity and soil organic matter are influence to connectivity of network of healthy soil, whereas degree of connectivity fungal networks of disease soil relatively low, and ammonium nitrogen and electrical conductivity were related to the connectivity of this network.

Concept 3: Differential networks focus on the changes. Challenging to network models by comparing networks across multiple environment is valuable. Networks can response differently under various environments or with external signals. They can be more simplify by focusing on key components and capturing only the essential components differently responding between environments which they play a key role in the modeled response (?, ?). Networks are examined by adding or depleting some variables. This allows predicting interactions or components that change following the changed structure of networks.

Data-driven computation, In silico network inference, can generate network very similar to those identified using experiment technique, which it can be used to identify the additional components interacting with the altered nodes, qualify and qualify the network after encountering perturbation. ? (?) showed clearly evidence that this computational prediction with experiment data can obtain novel opportunities for plant defense network modeling.

Summary

This literature has presented a brief introduction to the concepts and methods of network analysis. It has attempted to position network analysis as both a unique perspective and unique methodology with respect to analytic approaches and methods commonly used in plant pathology. Building on this foundation the literature ten identified two branches of plant pathology which network analysis has been identified as particularly useful, and examine theses. I concluded that network analysis has the potential to advance and operationalize certain aspects of plant disease epidemiology which mostly dealing with multidimensional data. In particular, network analysis has potential to help plant pathologists visualize, measure and document sources of epidemiological phenomena and to identify specific components contributing these situations.

Having developed a basic understanding of network analysis from both methodological and theoretical perspectives, the literature proceeds to review network concepts and methods in general. Roughly, Networks have four types, social network, information, technology network, and biological network. Even though, four types of networks are described and applied in different context. They share a common empirical focus on relational structure and a similar set of mathematical analyses. Node or vertex, link or edge. Network models are cable of presenting unique values, which are traditional approaches can not present.

Network concepts and methods are mainly found in the literature on two broad studies particularly in plant disease application. On the issue of understating plant disease spread, the literature suggests that further use of networks analysis concepts and methods enable us the holistic understand the flow of disease spread, and improve the implementation of plant disease policy. The literature of molecular plant pathology offers two challenge of network application. The fist challenges is to apply network to model large and complex biological dataset. Another challenge is consideration network structure to understand biological system. Emergent properties of network structure influences may be identified, measured and analyzed to yield better explanations of the experiments being observed.

While documented plant pathosystem studies using network analysis remain quite sparse, Using network analysis concepts and methods augment existing approaches and provide tools for exploring the relations dimensions, which has been widely acknowledged as influential but difficult to measure using traditional methods. This chapter claimed network analysis is ideally suited with high dimensional data, such as survey data, which is commonly used in the studies of plant pathosystem. With concepts and approaches giving generality and flexibility, networks can potentially model plant pathosystem as a subsystem of agroecosystem. Three concepts of network analysis are introduced to challenge the network application for system analysis in plant pathology. Addressing the challenges identified in this paper represent first steps in plant pathology to take advantage of these new opportunities.

CHAPTER III

MATERIALS AND METHODS

This research mainly examines relationships between the injuries caused by pests and diseases, production situations (e.g., rice varieties, crop establishments, fertilizer inputs, chemical applications), and rice yields using the data from surveys in irrigated lowland rice growing areas in South and South East Asia. I will develop and apply suitable methods of network analysis to characterize the patterns of co-occurrence of injuries and production situations. The resulting network of associations of injuries and production situations thus provides a starting point for further investigations of their relationships (i.e., comparison of networks from different production environments or examination of consequences of networks after imputation).

I propose three parts of network analysis of rice crop health survey. In the following, I present three distinct network analysis approaches: single-network analysis, differential network analysis, and dynamic network analysis. The three approaches answer different questions.

In the first part of network analysis, I apply single-network analysis in order to defines modules that can then be tested for validity with other data sets. Single-network analysis aims at identifying (a) patterns of interactions (modules) and (b) their key components (e.g., most connected variables) that are present in the data set.

The second part, differential network analysis, aims to uncover differences in the modules and connectivity between different data sets (e.g., dry season versus wet season). Each data set is then used to construct a network. Next, the networks are contrasted to find (1) non-preserved modules, (2) differentially occurred variables, and (3) differentially connected variables.

Dynamic network analysis, in third part, is applied for study changes of networks at least two different aspects of an evolving complex system. Here I vary yield gains, and obtain different yield data set in order to construct a dynamic network of yield varying behaviors. Similar to differential network analysis, dynamic networks focus on comparison of network structure, but it enable us to observe networks changing across successive yield gains.

Crop Health Survey Data

The crop health survey data were generated from surveys of farmers' fields in two seasons (wet and dry seasons) from 2009 to 2015 at different production environments across South and South East in irrigated lowland rice growing areas (West Java, Indonesia; Mekong River Delta and Red River Delta, Vietnam; Tamil Nadu and Odisha, India; and Suphanburi, Thailand) . They were conducted using the same protocol (?, ?).

The survey data consist of measures of multiple variables with different types of value. Data were divided each sample into three sets of variables, production situation set, injuries and disease set, and yield. **cropping practice set** are simplified, which collected with many type of data. For example, types of rice varieties (traditional varieties, modern varieties, and hybrid rice), crop establishments (direct seeded, transplanted rice) were collected in categorical data, pesticide (molluscicide, herbicide, insecticide and fungicide) uses were collected discretized data, and accumulated organic, chemical fertilizers were collected in continuous data. **Injuries and diseases set** composed of specific signs caused by pests or pathogens (i.e., whitehead, brown spot). They are collected percent of incidence of injury at two rice stages. Two types of injury indices were used areas under progress curves or maximum level of injuries or disease incidence depending on the nature of the injury. The time-dependent information on injuries was thus synthesized and compacted over time.

Samples composed of incomplete data were removed. These were encoded as a matrix in which each row represented a surveyed field in a specific location, year and season. Each column represented a collected variables.

Single Network Development

In the case of single network analysis, one use single network for modeling the relationship of cropping practice set, and injuries and disease sets. In the following, I describe a typical single-network analysis for finding the patterns of relationships. While a single network is the focus, it does not imply that only a single data set is used. Instead, appropriately similar multiple data sets can be used to validate the robustness of module definition and connectivity.

In the following, we provide an overview of single-network analysis strategy, which is depicted in Fig. 1: (a) process data preparation. (b) Calculate correlation coefficients (Pearson, Spearman, or Kendall). Estimate P-values for all coefficients. Next, determine threshold values for the resulting correlation coefficients and *P* values, storing results in adjacency matrices for the construction of networks. (c) Construct network and analyze network for graph-theoretic properties and infer biological meanings and integrate the network of input data and output data. (d) yield- related variables are used to prioritize variables within crop health data

STEP 1. *Evaluation of pairwise relationship association methods for network construction*

This research aims to construct networks visualizing the associations of injuries from pests and diseases with production situations. The rules defining edges of such networks is to present a sufficient level of ’association’ between certain attributes of the two nodes. I thus choose correlation measurements to construct an association network based on them.

To identify the most appropriate method for constructing a network based on correlation measurements, I select four correlation based measures, Pearson’s correlation, Spearman’s rank correlation, Kendall’s correlation, Biweight midcorrealtion. The cor.test function of R (?, ?) is applied for generating a correlation matrix, which describes the pairwise associations between variable in the context of the crop health survey data. This function allow users to select type of correlation measures to perform such as Pearson’s correlation, Spearman’s rank correlation and Kandell’s rank correlation. ”bicor” function of WGCNA package (?, ?) in R is applied for computing biweight midcorrelation matrix. A corresponding correlation matrix describes pairwise associations between variables is create.

STEP 2. *Comparison of threshold selection methods correlation matrix of survey data*

When a correlation matrix was create, next is to construct the correlation based network from the correlation matrix. However, the matrix is required the removal of spurious relationships by using threshold. Threshold is a value used for screening the correlations, if correlations below a threshold value, or close to zero, will be less meaningful, then will not be shown in network graphs. Threshold selection is statistically based, which can be obtained in two ways (?, ?)(1)

Determine *P* values for all pairwise comparison and adjust them for multiple hypotheses testing (e.g., Bonferroni or local false discovery rate. (2) Obtain a threshold value that guarantees a pre-specified false-discovery rate.

To select the right threshold value

STEP 3. *Evaluation of the network inference with prior knowledge of biological literature*

Although we can opt for a method based on its principle of statistical operation without paying attention to the biological models in a given data set, this may not lead to a coordination network that will reveal biological knowledge.

There is no statistical method that is suitable for all of them. Identification of the most efficient method for knowledge discovery of a specific biological process demands concrete prediagnostic analyses. Based on our study and our empirical knowledge, we would suggest the following procedure for identifying the most appropriate gene association method for a specific biological theme in a given data set: (1) Evaluate the prior knowledge of biological processes of one's interest, and select a few known genes involved in these processes; (2) Use the R codes from this study to perform a genome-wide coexpression analysis to obtain the top 100 or 500 genes that are most closely associated to the selected known genes; (3) Perform an evaluation of these 100 or 500 genes by examining which methods can associate the more functionally relevant genes to the selected genes. This can be achieved by examining gene annotation or performing GO term enrichment analysis: and (4) Choose the best method for the data. However, if prior knowledge of biological theme of one's interest is lacking, we suggest the most stable gene association method.

I

Differential Network Analysis

Networks allow one to look at components contributing to rice yield losses in systematic perspective. For instance, how are whitehead injuries related to type of crop establishment, or might leaf blast be related to insect pest injuries, or if a farmer used a direct seeding method, how are the risk of insect pests and other yield reducing factors related to this crop establishment method. This type of understanding is important, because if we can predict the key pests, then it is possible recommend a suitable pest management strategy for a given situation. It is anticipated that interesting connections between the individual inputs defined in production situation and single variable of injury profiles might appear through a network that have yet to be seen by more conventional approaches. I propose to collect the two types of profiling data; one is input profiling data, and another is output profiling data. The construction of networks from crop health survey data are illustrated in Figure **??**.

Dynamic Network Analysis

Analyzing the Structure of Network Models

Once networks are constructed, several indices can be computed that convey information about network structure. Structural properties of networks can be used for the interpretation of datasets and for generating hypotheses. Two types of structure are important. First, typically one is interested in the global structure of the network (random networks, small networks, scale-free networks) (?, ?, ?). Second, one may be interested in local patterns, which are characteristics of each node. For example, clustering of nodes and/or edges in a network can identify groups of nodes with similar properties, and these are referred to "modules" or "communities" (?, ?, ?).

For deep insights, comparing the networks by using some key topological properties of network are usefully conducted. Degree and degree distribution of a network is a simple property to extract from network models, which are the number of connections of each node and the frequency distribution of the number of connections per node, respectively. Cluster coefficient is the other measure, which is the value that is able to indicate whether the entities in network form cluster or group within network structure. ? (?, ?, ?) are recommended references for descriptions of the network properties as well as the formal calculation of these measures.

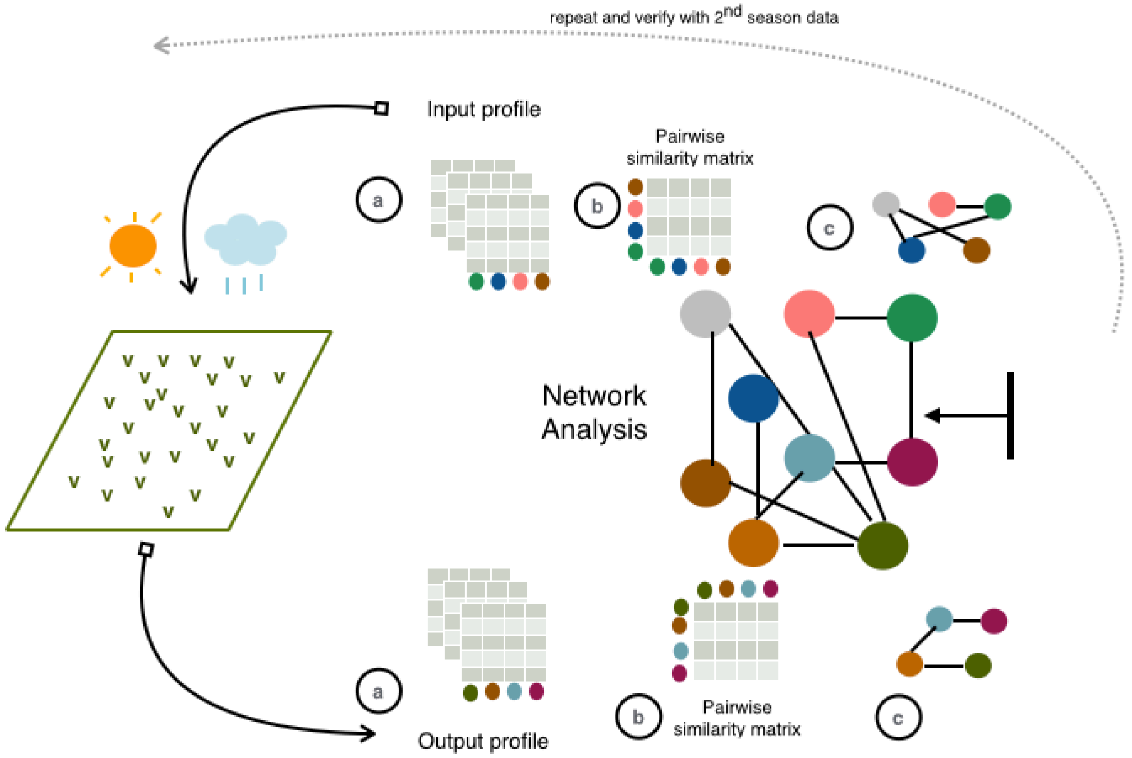
repeat and verify with 2nd season data
Input profile
a
Pairwise similarity matrix
b
c
Network Analysis
a
Output profile
b
Pairwise similarity matrix
c

Figure III.1: Proposed pipeline for network construction. (a) Collect the input profiling data and output profiling data from different samples and different locations. (b) Calculate correlation coefficients (Pearson, Spearman, or Kendall). Estimate P-values for all coefficients. Next, determine threshold values for the resulting correlation coefficients and *P* values, storing results in adjacency matrices for the construction of networks. (c) Construct network and analyze network for graph-theoretic properties and infer biological meanings and integrate the network of input data and output data. (d) Repeat analysis for a second season to verify the network model.

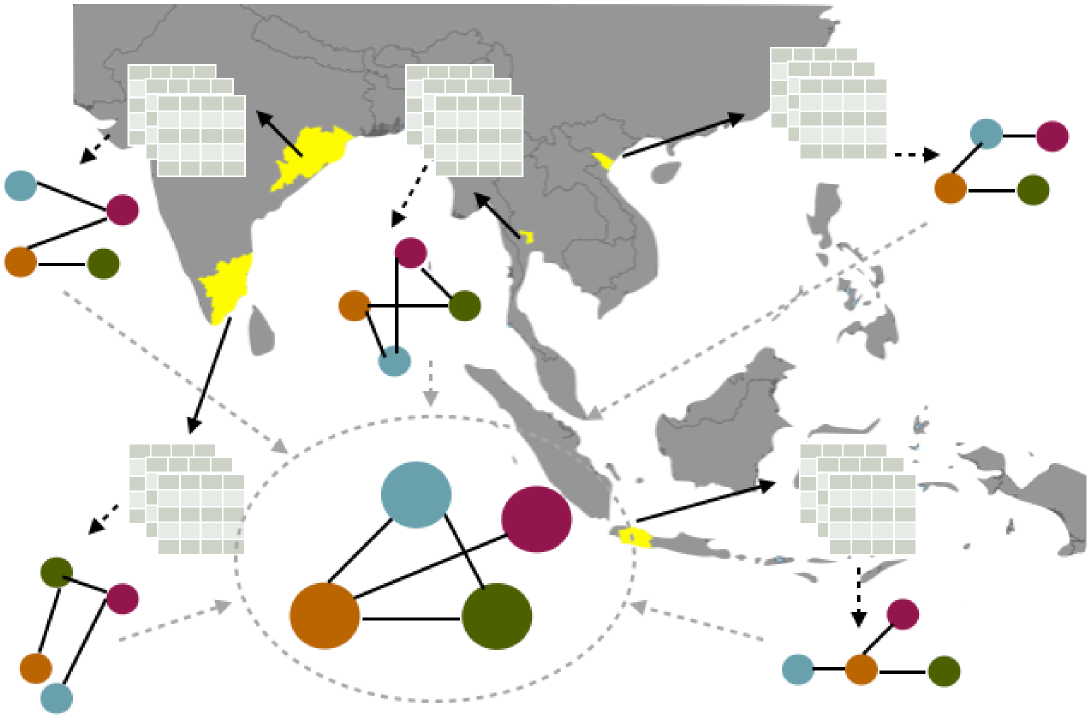

Figure III.2: Network comparison: Network models constructed from survey datasets of different geographic locations are compared by determining their properties. Networks will express the conserved domains within their structure. A merged representation of the two networks being compared is also proposed as a holistic network of rice ecosystem in South and South East Asia.