

**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**

# CHAPTER I

## INTRODUCTION

Pests and diseases to global rice production are significant yield reducing factors. ? (?) estimated that rice pests potentially caused losses around 37 percent of global rice production. Additionally, future rice production will need to grow by 2.4 percent per year in order to meet the demands of a growing population (?, ?). Addressing these yield-reducing factors is essential for food security not only in rice consuming societies, but also for other societies globally.

Rice is predominantly grown in Asia. Thirty-one percent of the rice harvested globally comes from Southeast Asia (?, ?). The highest levels of productivity are found in irrigated areas, the most intensified rice production system. Farmers can grow more than one rice crop per year here. Approximately 45 percent of the rice growing area in Southeast Asia is irrigated, with the largest areas being found in Indonesia, Viet Nam, Philippines and Thailand (?, ?). In South Asia, major rice-growing countries are India and Bangladesh. India has the largest rice growing area (approximately 43 million hectares) in the world and contributes 25 percent of global rice production.

Nowadays, developing the strategies of pest and disease management takes into account sustainability, production efficiency, and environmental protection (?, ?). To achieve this, in-

teractions between pests and human activities must be studied. A survey may provide the necessary data and adequate methods for analyzing survey data can produce preliminary information on their behaviors including major interactions (?). (?) concluded that the observed injury profiles (i.e., the combination of disease and pest injury that may occur in a given farmer's field) were strongly dependent on production situation. The authors discussed that pest management strategies should be developed according to the patterns of cropping practices, and production situations. However, interactions among pests, cropping practices, and environments are difficult to elucidate. Moreover, analytical techniques are found limitations in studying and revealing the differential interactions under different locations or over time, which they are important to design the strategies of pest management.

Network analysis provides a promising tool for revealing the interactions among entities within a complex system. It has been applied for many branches of science (e.g., social science, computer science, biology). A network model is an abstract model composed of a set of nodes or vertices and a set of edges, links or ties connected to the nodes. Nodes usually represent entities and the edges represent their relations. For example, an ecological network of a food web presents nodes as species (S<sub>1</sub>, S<sub>2</sub>) and edges as ecological relationships, or consider a social network of students in the school present where nodes are students and edges are friendships (S<sub>1</sub>, S<sub>2</sub>).

## **OBJECTIVES**

My overall objective is to develop network approaches and apply them to analyze crop health survey data. My first objective for this research is to develop the network model based on crop health survey data and characterize relationships among components of injuries and cropping practices. My second objective is to compare the differential relationships of network models

under different seasons or locations. The third objective is to apply network analysis to compare differential interactions in networks different levels of estimated actual yield from low to high.

Once network models based on crop health survey data are constructed, they can be helpful for plant health authorities and the people who related to crop protection especially for rice. The models support them 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 questions in the context of plant pathology.

This chapter contains four sections to thoroughly review of network analysis and its applications. In the first sections, I introduce a brief overview of the concepts and methods of network analysis, and I then discuss the unique values of network analysis that are not found in other approaches. In the third section I focus on the application of network analysis to current applications of plant pathology research, in particular, plant disease epidemiology and molecular plant pathology, for which network analysis has been broadly applied, and increasingly documented.

Network analysis 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 mechanisms and network models revealing spatial distribution of plant disease through trade networks were reviewed by ? (?) and ? (?), respectively.

## **Part I: Network Analysis**

### **Introduction to network analysis**

Network analysis is used 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 has 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.

Network analysis can be a powerful tool to study plant disease. ? (? , ? , ?) reviewed the applications of network analysis in botanical epidemiology and molecular plant pathology. Network models were applied to reveal the dynamics of the disease spread, or the plant defense mechanism during plant-pathogen interactions.



## Concepts, principles, and methods of network analysis

A network represents relationships between elements of interest, which are defined by links (edges) among nodes (vertices). Nodes can be units of interest or study, 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 (e.g, money, disease) and non-material (e.g., 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 goals of the representation and analysis (? , ?). Flow models, so-called directed graphs, view the network as a system of pathways along which somethings move such as transportation networks (e.g., of highways, railways and airlines) or communication networks. Since flows have directions, from an origin to a destination, the processes of movements are interesting to being modeled as a network model. Analysis of flow models can identify, which nodes in the network are more active, or which ones are more important connectors. ? (? , ?) applied such network models to study plant disease spread. Architectural models, so-called undirected graphs, are mainly used to determine the structure of the network, seeking to discern whether specific structures lead to similar outcomes or whether nodes 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 networks of soil microbial interactions. Network models can describe 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 an 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). A network link need not imply positive or cooperative interaction; they can also be a 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 (1, 2). 1 (1, 2, 3) are recommended references for descriptions of the theory and uses, as well as the formal calculation of these measures.

## Part II: The unique values of network analysis

There are four key points that will help to understand network analysis 1) how it differs from traditional approaches to social science research; 2) how it relates to those traditional approaches; 3) how networks are constructed, manipulated and measured; and 4) what value network analysis offers beyond traditional approaches.

The first point of network analysis is that there are two types of data resending in the network graphs; technical and rational data. The first is data characterizing the actors or variables being studied referring to attributes. Attributes describe characteristics of individual actors or variables, for example their race, income or physical location, and are the primary variables considered in traditional approaches. The second type of data is relational data, that is, data about the relationships between individual nodes. For example, ? (?) represented the network model of cooperation among lawyers in three law firms, through the exchange of various type of resources among them. This model consisted of over 70 lawyers in three different offices in three different cities. Rational data reflected to resources exchange, and additional attribute information were included type of practice, genders, and seniority of each lawyers. was recorded for each lawyers , including type of practice, genders, and seniority.

Relationships are also referred to as edges (links) in network analysis. Edges cannot be attributed to any single actor. Rather, edges only exist between nodes. This leads to the second point about network analysis that it requires a different conceptual approach. Because edges only exist between nodes, it is useful to think of edges existing in a separate dimension from nodes, who are anchored in physical space. This dimension is sometimes referred to as relational space. To visualize the difference, think of someone far away with whom you correspond regularly, say using a phone, email, or Facebook. Even though the two of you are not physically close, you have a strong relationship. The two of you are distant in physical space but close

in relational space. This notion of relational space is in part what means when he refers to the space of flows as something distinct from the space of places (?, ?).

The third point that distinguishes network analysis from other approaches is it involves different methods of analysis. Because traditional research methods consider variable attributes in a wide variety of statistical analyses such as measures of center (e.g., mean, median, etc) and dispersion (e.g., standard deviation, range etc.), these methods are sometimes referred to as variable analysis, whereas, network analysis models relational data and to measure various characteristics of network structure. For example, for lawyers data (?, ?), it is natural to ask to what extent two lawyers that both work with third lawyer are likely to with each other as well. This notion corresponds to the social network concept of transitivity and can be captured numerically through an enumeration of proportion of vertex triples that form triangles, so-called cluster coefficient.

The idea that network structure may be correlated with variable attributes and behaviors is the fourth point to consider in comparing network analysis to other approaches. In network analysis, the arrangement of the network in relational space is basically correlated with the behavior and attributes of those variables. For example, in the network created by ? (?) lawyers of the same firm may share similar attributes such as office location or department, and lawyers in similar roles within that network may share similar behaviors. Basically, conventional approaches measure various attributes of variable (nodes in a network) and attempts to discern something about the relationships between actors (edges in a network) based on those attributes. When the network structure is simple and the differences in node attributes are clear, the conventional analytic approach is sufficient. However when relationships are complex or node attributes are more nuanced, clear answers using conventional analysis may prove elusive. As a result, network analysis offers a tool to help researchers visualize the large network and disentangle

some of the relational complexities within the network, just as cluster analysis and multivariate analysis help research disentangle the complex data.

## **Part II: Networks and Plant Pathology**

Recently, a broad expansion of applications of network analysis has occurred across many disciplines over the past decade, and several researchers have evaluated network analysis as a promising tool to study a complex system. Plant pathologists also have used network analysis for their research. ? (? , ? , ?) supported that network analysis can be fruitful models in many applications relevant to plant pathology because of its generality and flexibility. For example, the network of main fresh cut flowers movements among European countries was determined the likelihood of introduction of new pathogens and other organisms associated with plants (? , ?), and plant-pathogen interaction network models were applied to present plant defense mechanisms (? , ?).

The development of network analysis challenges conventional approaches to uncover rational complexities of plant pathology studies. Two fields of research relevant to plant pathology presented particularly strong growth and proved 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 investigates questions related to biological networks.

### **Using Network analysis to understand plant disease spread**

Network analysis challenges conventional approaches of studies in plant disease epidemiology. Network approaches for spread of pathogens through trade network typify analytical risk assessment. For example, a structure analysis of network modeled the process of plant disease

spreads.

### ***Network models of epidemic development***

The idea of plant epidemics is that the probability of infection embedded in the connection or the contact patterns between susceptible/infected plants, and it forms as the networks. ? (?) showed a network model of epidemic development (susceptible-infected-susceptible model) in a directed network. In the network model, vertices were represented plant, and their attributes were presented the infectious status (healthy or infected). The epidemic is started at a single node, then nodes with a connection from the starting infected node will be infected at the next time step with a certain probability of transmission. In turn, already infected nodes will be infected 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 between infected nodes and susceptible nodes over times. Similarly, the probability of infection persistence is the same for infected nodes in a certain network replicate. For each network structure, the two probabilities of persistence and transmission define an epidemic threshold, which is independent of the starting node of the epidemic. This epidemiological model does not result in either susceptible or infected nodes, as nodes will have a infection status along a continuum. Key quantities for epidemiological dynamics in network were reviewed in ? (?).

### ***Analysis of plant trade network***

*Phytophthora ramorum* epidemic networks in the horticultural trade is an example of application of network models in the study of plant disease spreads (?, ?). Simulations of spread of *P. ramorum* in different network structure was found that epidemic threshold, the boundary between a no epidemic an epidemic outcome, is significantly lower for scale-free network, a network is dominated by a small number of nodes with many connections, compared to local, random and small-world network structure. Modeling suggested that was possible to control an

epidemic by changing the structure of network, without having to decrease the probability of infection persistence at a nursery site and/or of transmission between sites.

Regardless of the network structure and connectivity level, epidemic threshold is negatively correlated to the correlation coefficient between link in and out nodes, ( $\rho$ ,  $\rho$ ). In presence of high-connected nodes, the most effective way to control disease spread is to move from two-way to a one-way network, i.e. from network where overall there is positive correlation among links-in and -out to one where the correlation is negative. In practice this would mean that a nursery network would be dominated by major node, which receive plant materials from many production sites but supply a relatively few retail sites, or by major nodes which received plant materials from a few production sites but supply many retail sites. The scenario where there are major nodes which both receive plant material from many production sites and supply many retail sites is the most problematic control of this control is target towards such hubs. *P. ramorum*, these epidemic size would be the number of nurseries/retail centers with more than a certain proportion of plants infected, or the overall number of infected plant in all nurseries/retail centers. Simulations showed that the number of equilibrium. This correlation increase with connectivity level for all the structures investigated and underlines the importance of targeted control towards node with more connections than others ( $\rho$ ,  $\rho$ ).

The last point, the modeling of disease spread in small-size directed networks showed that increasing the proportion of wholesalers (i.e., traders without a preponderance of incoming or outgoing links) tends to decrease the epidemic threshold in no-scale free networks (local, random, and small-world ones) The opposite result is obtained for the proportions of producers and retailers. Scale free network appear instead to be immune to changes in these hierarchical categories, as the epidemic threshold in this case is governed by the presence of hub rather than by the features of the majority of nodes in the network.

### ***Network models to design strategies of plant disease management***

Due to globalization, increases in trade among countries affect to plant health. Network models presenting flows plant disease epidemics through plant trade network and develop strategies of plant disease management. Hubs or highly connected nodes represent locations or countries, where import and export plants or plant parts can be found in these network models. Because they take into account the likelihood of pathogen actually infecting along particular links following the network concepts, nodes are obviously targets for reduce disease flows in the scenarios that network model presented. Strategies for disease management may aim to remove them from the network or limit their geographic connectedness to try to alter the structure so as to slow and limit pathogen spread(?, ?). Alternatively, strategies may focus on prevention of disease spreads by throughout monitoring on hubs and highly connected nodes. For trade networks, ? (?) suggested that it would be sensible to place quarantine efforts on hubs or on connections between major hubs.

However, wherever effort is placed, an 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 health 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 (?, ?, ?, ?).

Additionally, ? (?) showed the good examples, which are co-occurrence network of the *Phytophthora ramorum* infected plant genera different environments. The networks may be



helpful in identifying host taxa playing a important role in spreading a certain disease in the seminatural environment, in crop plants, and plants in the trade. 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 approach, it was clear that the pathogen was introduced originally from nurseries, which *P. ramorum* populations in nurseries are genetically ancestral to all Californian forest populations.

## **Using Network analysis to understand molecular plant pathology**

Within the fields of molecular plant pathology, potential applications of network analysis include analysis of a protein's or gene's functions and their interactions in order to understand mechanisms of plant-pathogen interactions.

### ***Presenting biological data with network model***

Networks are used in different contexts as ways to represent relationships between entities, such as interactions between genes, proteins or metabolites. ? (? , ?) gave the example of a network model built from gene-for-gene relationships between rice and various avirulence genes of the pathogen *Xanthomonas oryzae* pv. *oryzae* causing bacterial leaf blight of rice. Nodes were represented isogenic lines of rice, and weighted edges reflected the number of shared genes with high resistance (with respect to avirulence genes) in the two isogenic lines of rice. For a plant breeder, this graph can help in identifying particularly promising genes for developing a variety that is resistant to pathogens.

***Network analysis to study biological systems*** Network analysis offers tools to visual the myriad information and analyze the complex relationships. To better understand the collective impact of genes on complex traits and determine what governs their organization, biologists are most likely to apply gene co-expression networks (? , ?). Co-expression networks most com-

monly use the Pearson correlation coefficient to establish linear pairwise correlations between gene pairs in an adjacency matrix. Another associative metric that can be used is the Spearman correlation coefficient which enables nonlinear correlations between genes to be uncovered (?, ?, ?). Once a coexpression network has been generated, identifying modules by clustering can help extract biological meaning from the network. Uncharacterized genes within such a module can be candidates for participating in the same process. Similarly, genes directly connected to (or co-expressed with) known central regulators of a developmental process are candidates within this co-functional framework.

? (?) used coexpression network inference to investigate plant immunity. The network modeled from transcriptome data sets of citrus infected with *Candidatus Liberibacter asiaticus* bacterium. This network revealed contained hub genes (genes may have similar functions), potentially key components of defense mechanisms, and novel genes that are responsible to defense mechanisms.

The main use of coexpression networks with large collections of static expression data is gene discovery. However, many biologists have attempted to construct differential networks to measure differences in connectivity patterns from datasets with different conditions or targeted experiments (control and treatment groups) to measure differences in connectivity patterns (?, ?). ? (?) constructed the networks of soil fungal communities. The fungal networks represented two different conditions, yield-invigorating and yield-debilitating soils, under prolonged potato monoculture were compared. The result showed that *Sordariales* and *Hypocreales* were the major affected phylogenetic groups. Network analysis enabled to identify the key elements and the patterns of relationships of ecological communities.

## Summary

This literature review presented a brief introduction of network analysis and concise concepts and methods. It also attempted to point out unique aspects and values that could not find in traditional approaches. To convince that network analysis can be applied in plant pathology, literature review showed the applications of network analysis to studies related to plant disease.

This literature review presented a brief Introduction to the concepts and methods of network analysis. 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. It attempted to position network analysis as both a unique perspective and unique methodology.

Network concepts and methods were applied to two broad studies particularly of plant disease. Network analysis was utilized to model plant disease spreads and to analyze the structure of models. The results enabled us to understand the directions and processes of disease spread. Additional, they predict how, and improve the implementation of plant disease policy. In molecular plant pathology the literature review showed two challenges of network applications. The first challenge is to apply networks to model large and complex biological datasets. 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 number of documented plant pathological studies using network analysis are sparse, the literature presented in this review showed a clear and compelling case for plant pathologists to expand the understanding of and use network concepts and methods. Network analysis

concepts and methods augment existing approaches and provide tools for exploring complex relationships, which have been widely acknowledged as influential but difficult to measure using traditional methods. With concepts and approaches giving generality and flexibility, networks can potentially model survey data.

# CHAPTER III

## MATERIALS AND METHODS

This research will use data collected from surveys of irrigated lowland rice growing areas in South and South East Asia to examine relationships between the injuries caused by pests and diseases, cropping practices (*e.g.*, rice variety grown, crop establishment method, fertilizer and chemicals applied), and rice yields. Their relationships will be constructed and analyzed through network analysis. I will develop and apply suitable methods of network analysis to characterize the associations of injuries and cropping practices. The resulting network of associations of injuries and cropping practices will thus provide a starting point for further investigations of their relationships (*i.e.*, comparison of networks under different production environments or examination of networks at different levels of yield gains).

Network analysis of rice crop health survey data is divided into three parts. In the following, I present three distinct network analysis approaches: single-network analysis, differential network analysis, and dynamic network analysis. These three approaches will answer different questions. I will apply single-network analysis to the data from all fields surveyed for identifying patterns of interactions between injuries and cropping practices and key components (*e.g.*, most connected variables). Second, differential network analysis will aim to uncover similari-

ties and differences of networks constructed from the different data sets (*e.g.*, dry season versus wet season). Dynamic network analysis, the third type, will be applied to study how networks changed under at least two different aspects of an evolving complex system. Here, I will focus on dynamic networks by dividing farms into different levels of yield attained.

## **Crop health survey data**

Crop health survey data were collected through surveys of farmers' fields from 2009 to 2015 for both wet and dry seasons in different production environments across South and South East Asia representing irrigated lowland rice growing areas West Java, Indonesia; Mekong River Delta and Red River Delta, Vietnam; Tamil Nadu and Odisha, India; and Suphanburi, Thailand. The survey protocol described in the IRRI publication, "A SURVEY PORTFOLIO TO CHARACTERIZE YIELD-REDUCING FACTORS IN RICE", (, ) was used for data collection. The variables collected included environmental attributes, patterns of cropping practices, crop growth measurement and crop management status assessments, measurements of levels of injuries caused by pests, and direct measurements of actual yields from crop cuts.

The survey data collected using the methodology described above consisted of measures of multiple variables. These variables can be classified into three groups: cropping practices, injuries, and actual yield measurements.

Injury data were gathered on diseases, insects and weeds observed at two development stages of the growing crop: active tillering and active ripening. While injuries due to diseases and insects were specific to species (or species groups), information on weed infestation was the area covered by any weed species, either above or below the crop canopy. Information pertaining to injuries was collected in the form of number of injured organs (tillers, leaves, and panicles), which later was made relative to the corresponding total number of organs present in

the sampling units (10 hills per field for transplanted rice crops). As for weed infestation, the proportion of soil area covered at two levels of the crop canopy (below or above it) was assessed in three points in the field of 1 m<sup>2</sup> each. For this purpose, two types of injury indices were used: area under injury progress curves or maximum level at any of the two observations, depending on the nature of the injury. The area under injury progress curve (AUIPC) was calculated by the mid- point method using the following equation:  $AUIPC = \sum \frac{1}{(x_i + X_{i-1})(T_i - T_{i-1})}$ .

Cropping practice data included information on the type of rice variety (traditional variety, modern variety, or hybrid rice), crop establishment methods used (direct seeded or transplanted rice) were collected as categorical data, pesticide usage (molluscicide, herbicide, insecticide and fungicide) were collected discretized data, and accumulated organic, synthetic fertilizers were collected as continuous data. Injuries and diseases composed of specific signs caused by pests or pathogens (*i.e.*, whitehead, brown spot) were collected as incidence of injury at two rice stages (active tillering and active ripening stage).

## Single network analysis of crop health survey data

### Introduction

Single-network analysis is used for analyzing a network based on a single data set. It aims to model patterns of relationships between entities to a network, thereby determining its topological properties such as node degree, degree distribution, average path length, clustering coefficient, modularity index ( $Q$ ,  $Q$ ) to describe the patterns in the network model. These structural properties offer the potential to explore relationships within complex data sets. Additionally, it is used to identify key drivers, which are key factors or clusters.

Correlation-based networks are applied to reveal the patterns of interactions or relationships of biological system. Their edges are obtained using correlation-based measures removing

spurious relationships. Correlation based measures include linear based correlation (Pearson's correlation, partial correlation) and rank based correlation (Spearman's correlation, Kendall's correlation). Removal of spurious relationships is particularly important when one attempts to establish causal relationships between entities (?, ?).

In the case of single network analysis, a network model of cropping practices and injury profiles based on whole data set of surveys will be defined using a correlation as an undirected, weighted network. Nodes of the network corresponded to variables and edges between variables will be determined by their pairwise correlation.

A network model based on crop health survey data has not yet been reported. Additionally, correlation analysis and network analysis of these data has not been studied. Thus, correlation measures and network approaches should be evaluated.

According to ? (?), correlation based network construction has at least three important issues to be dealt with. First, there is the choice of pair-wise correlation measure to be used. Second, the level of statistical significance must be determined for removal of sparse relationships. Finally, the problem of multiple testing must be considered because there are large numbers of tests for pair-wise comparison.

## **Single network analysis network construction**

Cropping practices and injury profiles network are constructed by an adjacency matrix originating from the pair-wise correlations of the incidence of injuries (insect injuries and diseases), amount of fertilizer, the frequency of pesticide applied, type of rice varieties and crop establishment methods across different samples. Therefore, edges of the network correspond to the degree of correlation between two variables. A standard measurement of correlation between two variable  $x$  and  $y$ ,  $cor(x, y)$ , where values are between -1 to +1 depending on the level of



relationship.  $cor(x_i, x_j)$  is equal to -1 when there is a decreasing relationship between  $x$  and  $y$ , and +1 when there is a increasing relationship.

To identify the most suitable pairwise correlation methods, I will evaluate four correlation based measures, Pearson's correlation, Spearman's rank correlation, Kendall's correlation, and biweight midcorrelation. R was employed to compute pairwise correlation (Wang, 2014). The `cor.test()` function will be applied for generating a correlation matrix, which describes the pairwise correlations between variables. This function allows users to choose types of correlation measures to perform such as Pearson's correlation, Spearman's rank correlation and Kendall's rank correlation. To compute biweight midcorrelation, I will apply the `bicor()` function from the **WGCNA** package (Wang, 2014) in R.

When a correlation matrix is created, next is to construct the correlation based network. However,  $p$  values should be concerned. As with issues previously mentioned above,  $p$  values must be adjusted for multiple testing. Using a Benjamini-Hochberg adjustment or Bonferonni correction is recommended by Wang (2014). The `p.adjust()` function in R can calculate adjusted  $p$  values. These values are compared to a standard 0.05 significance level. The final correlation matrix contains pair-wise correlation coefficients, which have adjusted for  $p$  values lower than 0.05 significance level. Alternatively, they are adjusted through control of the false discovery rate. The R package **fdrtool** can be implemented for this method (Wang, 2014).

Once networks based on the four association metrics, Pearson coefficient, and Spearman coefficient, Kendall coefficient and Biweight midcorrelation, have been constructed and assessed basic properties, I will use the following R packages: `igraph` (Csardi & Nepusz, 2006); `qgraph` (Epskamp et al., 2012); `statnet` (Hahn et al., 2012) and `network` (Butts, 2008); and `sna` (Garg, 2012) to construct a network graph based on survey data, to decorate the network, and then analyze the network structure.

A network modeled from biological data may not reveal useful biological information, even

though it is created from a method based on its principle of statistical operation (?, ?). For correlation based networks edges correspond to correlation coefficients and also affect to the network structure and behaviors. Different correlation based measures contribute to networks with different structures, even though they are constructed from the same data set. I will investigate the efficiency of networks based on four correlation measures for revealing the associations between variables of surveys. The most suitable correlation measure is assumed that it can contribute the network being able to reveal the associations between variables based on the existing literatures or good understandings of biological relationships between variables that crop health surveys observed.

## Evaluating network properties

Once a network is constructed several indices can be computed to convey information about network structure. To evaluate the topological properties of both the interaction and the correlation based network, I will use the packages **igraph** (?, ?) and **qgraph** (?, ?) in R. I particularly are interested in properties potentially relevant for biological roles and functioning as previously hypothesized in other biological networks (?, ?, ?, ?).

1. Mean degree: the degree of a node counts the number of edges it has. The mean degree is calculated over all nodes in the network.
2. Degree distribution: the frequency of nodes vs. their (increasing) degree.
3. Average shortest path length: the shortest path between any two nodes is the single path with fewest edges between them. Alternative paths are feasible. The average shortest path length is the mean over all shortest paths between any two nodes in the network.
4. Mean clustering coefficient: a cluster of nodes is a triangle of nodes. The clustering

coefficient calculates the fraction of observed vs. possible triangles for each node. The mean is subsequently determined from all nodes in the network.

5. Betweenness centrality: the betweenness centrality of a node is equal to the number of shortest paths between any two nodes in the graph passing through that node. The mean is calculated from all nodes in the network.
6. Closeness centrality: the closeness centrality of a node is given by the average distance of this node to any other node. Again, the network-wide measure is an average over all nodes in the network.

? (? , ? , ? , ? ) are recommended references for descriptions of the network properties as well as the formal calculation of these measures.

## **Differential network analysis of crop heath survey data under different seasons and production environments**

Networks can respond differently under various environments or with external signals. They can be simplified by focusing on key components and capture only the essential components differently responding between environments which they play a key role in the modeled response (? , ?). Networks are examined by adding or deleting some variables. This allows predicting interactions or components that change following the changed structure of networks.

Differential network analysis is applied to identify and describe differences between two networks under different conditions. Differential networks are created from different data sets or subsetting(?) from one data set with different condition. These sub-networks might display different interactions from the single network based on the whole data set. The strongest inter-

actions in differential networks are not necessarily those that are strong in the single network. Conversely, the interactions are weak or removed from the differential network. If networks constructed from different environment conditions, the differential interactions between two networks are implied changes that they are a result of response to environmental conditions. Moreover, it can implied that environment conditions influence on the behaviors of pairs of nodes contributing to the differential interactions.

The objective of differential network analysis is to extract interactions from the original network that appear to be active under different conditions. I focus on two aspects of this analysis. The first is to compare how patterns of interactions of cropping practices and injury profile network change across multiple production environments. The second is to study how networks change with different cropping seasons (wet or dry). Each data set is used to construct the network. Next, these networks are contrasted to find a) non-preserved modules, b) differentially occurred variables, and c) differentially connected variables.

Based on the methods of single network construction mentioned above, differential networks will be constructed from the survey data with different groups of samples following the purpose of network construction. To analyze differential networks under different seasons, I will construct two networks; one will be constructed from dry season data, and another will be constructed from wet season data. Additionally I will determine differential networks under different production environments (locations).

For comparison with a standard differential network analysis, I applied **WGCNA** (Wang, 2013) and **dna** (Differential Network Analysis) package in R, which this package provide several function to analyze the differences of topologies of two networks (Wang, 2013).

These packages include preprocessing tools for simultaneously preparing a pair of networks for analysis, procedures for computing connectivity scores between pairs of variables based on

many available statistical techniques, and tools for handling modules of variables based on these scores. Also, procedures are provided for performing permutation tests based on these scores to determine if the connectivity of variables differs between the two networks, to determine if the connectivity of a particular set of important variables differs between the two networks, and to determine if the overall module structure differs between the two networks. Several built-in options are available for the types of scores and distances used in the testing procedures, and additionally, the procedures provide flexible methods that allow the user to define custom scores and distances. For example, the `test.modular.structure()` function is used to compare between the connectivity measures of each network (?, ?).

## **Dynamic network analysis of crop health survey data**

Networks can dynamically respond and adapt to the internal state and external signals (?, ?). As the internal state, backgrounds of nodes influence the structure and behavior of the network and give rise to significant differences 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.

Biological systems are highly dynamic. They must continuously respond to a external signals or the internal state of the system. The responses can be altered slowly or quickly over time (?, ?). Thus, realistically the corresponding biological networks evolve as well. It seems clear that if we are able to construct dynamic networks that enable us to see and understand the system and how their dynamic effects are changed over time, we will have a better understanding of the system. Some understandings previously have been obtained from studies of dynamics of large networks, for example, gene expression or metabolic fluxes network (?, ?).

Dynamic network analysis will be applied to study changes in networks with at least two

different aspects of an evolving complex system.

The main goal of dynamic network analysis moves away from single network analysis, which characterizes absolute properties of the system. It aims to concentrate on a specific dynamic systems response. Rather than answer what the key factors in the system, they answer what parts of the system are most affected by perturbation. Most commonly, dynamic networks are applied when the edges among a set of vertices or the sets of vertices itself are changing as a function of time. However, in the context of crop health surveys, I will apply dynamic networks to capture the yield-varying behaviors through different survey data sets are grouped by different levels of yield measurements.

My objectives are to construct and analyze networks of crop health survey data at different levels of farmers' yields. Surveys contain records of observed actual yields, various types of injuries and cropping practices. To generate the data set for constructing the dynamic network, I will group the data with successive levels of yields to obtain different yield data sets in order to construct a dynamic network of yield-varying behaviors. I then will employ **networkDynamic** (?, ?) and **ndtv** (?, ?) packages to generate yield-varying networks. The dynamic graphs will be characterized following (?, ?, ?). From the network based perspectives, the results will show the patterns of interactions between nodes how they changed when levels of yield decreased or increased.

# LITERATURE CITED

- Barabasi, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2), 101–113.
- Bender-deMoll, S. (2015). ndtv: Network dynamic temporal visualizations [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=ndtv> (R package version 0.6.1)
- Bilgin, C. C., & Yener, B. (2006). Dynamic network evolution: Models, clustering, anomaly detection. *IEEE Networks*.
- Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*. SAGE Publications Limited.
- Butts, C. T. (2015). network: Classes for relational data [Computer software manual]. Retrieved from [CRAN.R-project.org/package=network](http://CRAN.R-project.org/package=network) (R package version 1.12.0)
- Butts, C. T., Leslie-Cook, A., Krivitsky, P. N., & Bender-deMoll, S. (2014). networkdynamic: Dynamic extensions for network objects [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=networkDynamic> (R package version 0.7.1)
- Castells, M. (2001). *The rise of network society*. Oxford: Blackwell Publishers.
- Csardi, G., & Nepusz, T. (2010). Csardi: igraph: Network analysis and visualization. *R package version 05*.
- Dehnen-Schmutz, K., Holdenrieder, O., Jeger, M. J., & Pautasso, M. (2010). Structural change

- in the international horticultural industry: some implications for plant health. *Scientia Horticulturae*, 125(1), 1–15.
- Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., & Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinformatics*, 13, 113.
- Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, 48(4), 1–18. Retrieved from <http://www.jstatsoft.org/v48/i04/>
- FAO, O. (2012). *OECD-FAO Agricultural Outlook 2012–2021*. OECD Publishing and FAO Paris.
- Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Computational Biology*, 8(7).
- Freilich, S., Kreimer, A., Meilijson, I., Gophna, U., Sharan, R., & Ruppin, E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic acids research*, 38(12), 3857–3868.
- Gill, R., Datta, S., & Datta, S. (2014). dna: Differential network analysis [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=dna> (R package version 1.1-1)
- Guimera, R., & Amaral, L. A. N. (2005). Functional cartography of complex metabolic networks. *Nature*, 433(7028), 895–900.
- Horvath, S. (2011). *Weighted network analysis: Applications in genomics and systems biology*. Springer Science & Business Media.
- Ideker, T., & Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8(1). Retrieved from <http://dx.doi.org/10.1038/msb.2011.99> doi:



- Jeger, M. J., Pautasso, M., & Holdenrieder, O. (2007). Modelling disease spread and control in networks: implications for plant sciences. *New Pathologist*.
- Kasari, C., Locke, J., Gulsrud, A., & Rotheram-Fuller, E. (2011). Social networks and friendships at school: Comparing children with and without asd. *Journal of autism and developmental disorders*, 41(5), 533–544.
- Klaus, B., & Strimmer, K. (2015). fdrtool: Estimation of (local) false discovery rates and higher criticism [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=fdrtool> (R package version 1.2.14)
- Kolaczyk, E. D., & Csárdi, G. (2014). *Statistical analysis of network data with r* (Vol. 65). Springer.
- Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E., & Taylor, W. W. (2003). Compartments revealed in food-web structure. *Nature*, 426(6964), 282–285.
- Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559.
- Lazega, E., & Pattison, P. (1999). Multiplexity, generalized exchange and cooperation in organizations: a case study. *Soc. Network*, 21(1), 67–90.
- Lu, L., Yin, S., Liu, X., Zhang, W., Gu, T., Shen, Q., & Qiu, H. (2013). Fungal networks in yield-invigorating and -debilitating soils induced by prolonged potato monoculture. *Soil Biology and Biochemistry*, 65, 186–194.
- Mew, T. W., Leung, H., Savary, S., Vera Cruz, C. M., & Leach, J. E. (2004). Looking ahead in rice disease research and management. *Critical Reviews in Plant Sciences*, 23(2), 103–127.
- Moody, J. (2001). Race, school integration, and friendship segregation in america1. *American*

- Journal of Sociology*, 107(3), 679–716.
- Moslonka-Lefebvre, M., Finley, A., Dorigatti, I., Dehnen-Schmutz, K., Harwood, T., Jeger, M. J., Xu, X., Holdenrieder, O., & Pautasso, M. (2011). Networks in plant epidemiology: from genes to landscapes, countries, and continents. *Phytopathology*, 101(4), 392–403.
- Moslonka-Lefebvre, M., Mathieu Pautasso, M., & Jeger, M. J. (2009). Disease spread in small-size directed networks: epidemic threshold, correlation between links to and from nodes, and clustering. *Journal of Theoretical Biology*, 206, 402–411.
- Mutert, E., & Fairhurst, T. (2002). Developments in rice production in southeast asia. *Better Crops International*, 15, 12–17.
- Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*, 45(2), 167–256.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23), 8577–8582.
- Oerke, E. -C. (2006). Crop losses to pests. *The Journal of Agricultural Science*, 144(1), 31–43.
- Pastor-Satorras, R., & Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical review letters*, 86(14), 3200.
- Pautass, M., Moslonka-Lefebvre, M., & Jeger, M. J. (2010). The number of links to and from the starting node as a predictor of epidemic size in small-size directed networks. *Ecological Complexity*, 7, 424–432.
- Pe’er, D., & Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, 144(6), 864–873.
- Proulx, S., Promislow, D., & Phillips, P. (2005). Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, 20(6), 345–353.
- Ray, D. K., Mueller, N. D., West, P. C., & Foley, J. A. (2013). Yield trends are insufficient to

- double global crop production by 2050. *PLoS ONE*, 8(6).
- Savary, S., & Castilla, N. (2009). A survey portfolio to characterize yield-reducing factors in rice. *IRRI Discussion Paper No 18*.
- Savary, S., Madden, L., Zadoks, J., & Klein-Gebbinck, H. (1995). Use of categorical information and correspondence analysis in plant disease epidemiology. *Advances in botanical research*, 21, 213–240.
- Savary, S., Willocquet, L., Elazegui, F. A., Teng, P. S., Van Du, P., Zhu, D., Tang, Q., Huang, S., Lin, X., & Singh, H. M. (2000a). Rice pest constraints in tropical Asia: characterization of injury profiles in relation to production situations. *Plant disease*, 84(3), 341–356.
- Shaw, M. W., & Pautasso, M. (2014). Networks and plant disease management: Concepts and applications. *Annual Review of Phytopathology*, 52(1), 477–493.
- Strogatz, S. H. (2001). Exploring complex networks. *Nature*.
- Team, R. C. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Toubiana, D., Fernie, A. R., Nikoloski, Z., & Fait, A. (2013). Network analysis: tackling complex data to study plant metabolism. *Trends in biotechnology*, 31(1), 29–36.
- Usadel, B., Obayashi, T., Mutwil, M., Giorgi, F. M., Bassel, G. W., Tanimoto, M., Chow, A., Steinhauser, D., Persson, S., & Provart, N. J. (2009). Co-expression tools for plant biology: opportunities for hypothesis generation and caveats. *Plant, cell & environment*, 32(12), 1633–1651.
- Windram, O., Penfold, C. A., & Denby, K. J. (2014). Network modeling to understand plant immunity. *Annual Review of Phytopathology*.
- Wu, X., Li, Y., Zou, L., & Chen, G. (2007). Gene-for-gene relationships between rice and diverse avrbs3/ptha avirulence genes in *Xanthomonas oryzae* pv. *oryzae*. *Plant Pathology*,

56(1), 26–34.

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., & Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nature communications*, 5.

Zheng, Z. L., & Zhao, Y. (2013). Transcriptome comparison and gene coexpression network analysis provide a systems view of citrus response to "Candidatus Liberibacter asiaticus" infection. *BMC Genomics*, 14, 27.