Identifying highly connected sites for risk-based surveillance and control of cucurbit downy mildew in the eastern United States (#93930)

First submission

Guidance from your Editor

Please submit by **23 Jan 2024** for the benefit of the authors (and your token reward) .

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Author notes
Have you read the author notes on the guidance page?

Raw data check
Review the raw data.

Image check
Check that figures and images have not been inappropriately manipulated.

If this article is published your review will be made public. You can choose whether to sign your review. If uploading a PDF please remove any identifiable information (if you want to remain anonymous).

Files
Download and review all files from the materials page.

13 Figure file(s)
4 Other file(s)

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult. I suggest you have a colleague who is proficient in English and familiar with the subject matter review your manuscript, or contact a professional editing service.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Identifying highly connected sites for risk-based surveillance and control of cucurbit downy mildew in the eastern United States

Awino M. E. Ojwang' [1], **Alun L Lloyd** [1], **Sharmodeep Bhattacharyya** [2], **Shirshendu Chatterjee** [3], **David H Gent** [4], **Peter S Ojiambo** [Corresp. 5]

[1] Biomathematics Graduate Program and Department of Mathematics, North Carolina State University, Raleigh, North Carolina, United States

[2] Department of Statistics, Oregon State University, Corvallis, OR, United States

[3] Department of Mathematics, City University of New York, City College, New York, NY, United States

[4] Agricultural Research Service, U.S. Department of Agriculture, Corvallis, OR, United States

[5] Center for Integrated Fungal Research, Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, North Carolina, United States

Corresponding Author: Peter S Ojiambo
Email address: pojiamb@ncsu.edu

Objective. Surveillance is critical for the rapid implementation of control measures for diseases caused by aerially dispersed plant pathogens, but such programs can be resource-intensive, especially for epidemics caused by long-distance dispersed pathogens. The current cucurbit downy mildew platform for monitoring, predicting and communicating the risk of disease spread in the United States is expensive to maintain. In this study, we focused on identifying sites critical for surveillance and treatment in an attempt to reduce disease monitoring costs and where control may be applied to mitigate the risk of disease spread.

Methods. Static networks were constructed based on the distance between fields, while dynamic networks were constructed based on the distance between fields and wind speed and direction, using epidemic data collected from 2008 to 2016. Three strategies were used to identify highly connected field sites. First, the probability of pathogen transmission between nodes and the probability of node infection were modeled over a discrete weekly time step within an epidemic year. Second, nodes identified as important were selectively removed from networks and the probability of node infection was recalculated in each epidemic year. Third, the recurring patterns of node infection were analyzed across epidemic years.

Results. Static networks exhibited scale-free properties where the node degree followed a power-law distribution. Betweenness centrality was the most useful metric for identifying important nodes within the networks that were associated with disease transmission and prediction. Based on betweenness centrality, field sites in Maryland, North Carolina, Ohio, South Carolina and Virginia were the most central in the disease network across epidemic years. Removing field sites identified as important limited the predicted risk of disease spread based on the dynamic network model.

Conclusions. Combining the dynamic network model and centrality metrics facilitated the identification of highly connected fields in the southeastern United States and the mid-Atlantic region. These highly connected sites may be used to inform surveillance and strategies for controlling cucurbit downy mildew in the eastern United States.

Identifying highly connected sites for risk-based surveillance and control of cucurbit downy mildew in the eastern United States

Awino M. E. Ojwang'[1], Alun L. Lloyd[1], Sharmodeep Bhattacharyya[2], Shirshendu Chatterjee[3], David H. Gent[4] and Peter S. Ojiambo[5]

[1] Biomathematics Graduate Program and Department of Mathematics, North Carolina State University, Raleigh, NC, USA

[2] Department of Statistics, Oregon State University, Corvallis, OR, USA

[3] Department of Mathematics, City University of New York, New York, NY, USA

[4] U.S. Department of Agriculture, Agricultural Research Service, Corvallis, OR, USA

[5] Center for Integrated Fungal Research, Department of Entomology and Plant Pathology, North Carolina State University, Raleigh, NC, USA

Corresponding author: P.S. Ojiambo, pojiamb@ncsu.edu

ABSTRACT

Objective. Surveillance is critical for the rapid implementation of control measures for diseases caused by aerially dispersed plant pathogens, but such programs can be resource-intensive, especially for epidemics caused by long-distance dispersed pathogens. The current cucurbit downy mildew platform for monitoring, predicting and communicating the risk of disease spread in the United States is expensive to maintain. In this study, we focused on identifying sites critical for surveillance and treatment in an attempt to reduce disease monitoring costs and where control may be applied to mitigate the risk of disease spread.

Methods. Static networks were constructed based on the distance between fields, while dynamic networks were constructed based on the distance between fields and wind speed and direction, using epidemic data collected from 2008 to 2016. Three strategies were used to identify highly connected field sites. First, the probability of pathogen transmission between nodes and the probability of node infection were modeled over a discrete weekly time step within an epidemic year. Second, nodes identified as important were selectively removed from networks and the probability of node infection was recalculated in each epidemic year. Third, the recurring patterns of node infection were analyzed across epidemic years.

Results. Static networks exhibited scale-free properties where the node degree followed a power-law distribution. Betweenness centrality was the most useful metric for identifying important nodes within the networks that were associated with disease transmission and prediction. Based on betweenness centrality, field sites in Maryland, North Carolina, Ohio, South Carolina and Virginia were the most central in the disease network across epidemic years. Removing field sites identified as important limited the predicted risk of disease spread based on the dynamic network model.

Conclusions. Combining the dynamic network model and centrality metrics facilitated the identification of highly connected fields in the southeastern United States and the mid-Atlantic region. These highly connected sites may be used to inform surveillance and strategies for controlling cucurbit downy mildew in the eastern United States.

Subjects: Computational Biology, Ecology, Epidemiology, Plant Science, Statistics

Keywords: Centrality measures, Disease monitoring, Infection frequency, Network analysis, Scale-free network

INTRODUCTION

Pathogen dispersal is a fundamental property in developing disease epidemics at different spatial scales that can range from local to the landscape level. The transmission of invasive plant pathogens and the spread of resultant epidemics influences essential ecosystem services, including biodiversity and food production in agricultural systems (*Brown & Hovmøller*, *2002*; *Crowl et al.*, 2008). Measures that involve containment and eradication programs can be implemented to reduce the potential impact of these epidemics. However, the planning and implementation of any specific measure requires an understanding of the mechanics of invasions and the ecological consequences, risks, and dynamics of disease spread. Such efforts can benefit greatly from epidemic records within a region as they enable an analysis of the overall structure of pathogen dispersal. Information from such analyses can help design control programs for disease epidemics and risk-based surveillance. For example, timely recording of animal movements was fundamental in the containment of the 2011 foot and mouth disease epidemic in the UK, for which retrospective analyses demonstrated that initial spread was influenced by the frequency of animal movement (*Ferguson, Donnelly & Anderson, 2001; Kao et al., 2006*).

One approach to understand pathogen dispersal and the spread of resultant epidemics is through network analysis, a method that is becoming increasingly popular but still has limited application in plant disease epidemiology (*Garrett et al., 2018; Xing et al., 2020*). Networks consist of 'nodes' and 'links', where nodes are the entities of interest (e.g., individual fields or observed sites of disease outbreak), while links connect nodes in various ways, for example, the potential of encounter or pathogen transmission between two nodes. Further, networks can be weighted with link weights that are proportional to the probability of transmission. Networks have been used to describe the spread of epidemics caused by aerially dispersed plant pathogens such

as *Podosphaera macularis* in hop (*Gent, Bhattacharyya & Ruiz, 2019*) and *Phakopsora pachyrhizi* in soybean (*Sutrave et al., 2012; Sanatkar et al., 2015*). The primary determinants in pathogen dispersal, such as source strength, location of host populations and relevant covariate information, can be formulated as a network spreading model (*Firester, Shtienberg & Blank, 2018*; *Garrett et al., 2018; Gent, Bhattacharyya & Ruiz, 2019; Sutrave et al., 2012*). Such models can combine static spatial components, such as field location, and dynamic components of an epidemic, such as wind, to infer the underlying contact structure of landscape connectivity (*With et al*., *1997*).

The choice of networks to be studied depends on several factors that include the disease of interest and specific questions on the network structure. The latter, in turn, influences the type of network measures to be used in the analysis of pathogen dispersal and disease spread. Static and dynamic networks are common in landscape connectivity analyses. Connections in a static network are fixed links, while connections in a dynamic network change over time. Both static and dynamic networks have been applied in plant disease epidemiology (*Sanatkar et al., 2015; Sutrave et al., 2012*). In dynamic networks, between-node distances, host availability, wind speed and wind direction, can be formulated as a susceptible-infected (SI) model to describe epidemic spread (*Sutrave et al., 2012*). Further, plant diseases display seasonal differences in the occurrence and intensity of epidemics. Thus, analysis of data from multiple epidemic years is useful in determining if there are recurring patterns that could inform monitoring or disease control measures. Highly connected nodes provide more effective surveillance and opportunities for more targeted control to reduce disease spread within the network. An open question still remains regarding which centrality measures are most important for identifying important nodes for surveillance and managing real-world networks (*Holme, 2017*). Due to inherent differences in pathogen dispersal

and disease spread mechanisms, centrality measures used to identify important nodes for surveillance may be specific to different pathosystems (*Holme*, *2018*).

A motivating plant disease example for network analysis to inform surveillance and disease control is cucurbit downy mildew (CDM). A resurgence of the disease occurred around the world in the last 20 years that fundamentally altered cucurbit production and disease management at multiple scales (*Holmes et al.*, *2015*; *Ojiambo et al.*, *2015*). The resurgence of CDM in Europe and the United States was attributed to the introduction of a new pathotype or species that was previously limited to East Asia (*Cohen et al.*, *2015*; *Thomas et al.*, *2017*). Fungicides are integral to CDM control due to the lack of cultivars with adequate resistance and in the absence of control, the disease can result in complete crop loss (*Holmes et al.*, *2015*). The disease is caused by an obligate pathogen, *Pseudoperonospora cubensis*, which exhibits significant long-distance dispersal (*Ojiambo* & *Holmes*, *2011*). In continental United States, *P. cubensis* overwinters below approximately 30-degree latitude in southern Florida and along the Gulf of Mexico on living hosts, and disease outbreaks in northern states rely on pathogen dispersal from the south (*Ojwang' et al.*, *2021*). In 2008, disease surveillance based on a series of sentinel and non-sentinel field sites was implemented as part of the CDM ipmPIPE (http://cdm.ipmpipe.org) surveillance system (*Ojiambo et al.*, *2011*). Based on the prediction framework developed by *Main et al.* (*2001*) and the sentinel site data, an integrated aerobiological model was developed to predict disease occurrence and progression in the eastern United States (*Neufeld et al.*, *2018*) to guide growers on when to apply their initial fungicide application. Surveys conducted in Georgia, Michigan, and North Carolina show that the forecasting system resulted in an average reduction of two to three fungicide applications compared to calendar-based application schedules. This reduction in fungicide applications translated to more than $6 million in savings for producers in these three states alone

annually (*Ojiambo et al*., *2011*). However, the disease surveillance system is expensive to maintain and thus, there is increasing interest in identifying locations that are critical for pathogen dispersal and disease spread within the region. The latter could facilitate a more targeted surveillance approach by directing the limited resources to locations that are more integral to disease spread and pathogen transmission within the region. These sentinel and non-sentinel sites have been instrumental in understanding the spatio-temporal spread of CDM (*Ojiambo & Holmes*, *2011*; *Ojiambo et al*., *2017*; *Ojwang' et al*., *2021*).

In this study, we specifically focus on centrality metrics (*Meghanathan & Lawrence*, *2016*) that are directly applicable for CDM surveillance and management to identify highly connected sites. The centrality measures are betweenness (BWC), closeness (CLC), degree (DGC) and eigenvector (EVC), that have previously been used in network analysis of aerially dispersed plant pathogens and have relevance in describing epidemic spread (*Andersen et al., 2019*). Our inference of the importance of the highly connected sites is limited to disease records from the existing structure of sentinel and non-sentinel sites within the region. The specific objectives of this study were to: i) determine a centrality measure that is most useful in the surveillance and control of CDM, ii) identify highly connected nodes that are critical for pathogen dispersal and spread of CDM and iii) establish how removal of highly connected nodes influences the spread and containment of CDM in the eastern United States.

MATERIALS AND METHODS

Data source

Records of CDM epidemics in the eastern United States from 2008 to 2016 were used in this study. The data were obtained from the CDM ipmPIPE database (http://cdm.ipmpipe.org) that tracks

reports of disease occurrence in the United States (*Ojiambo et al., 2011*). Epidemic records in the system include reports from a network of regularly monitored sites (sentinel plots) and voluntary reports (non-sentinel sites) submitted by commercial growers, agricultural researchers and the public. Sentinel sites were strategically placed within specific states and planted with different cucurbit host types for monitoring CDM occurrence. Sentinel sites were located at research facilities or commercial fields with standard dimensions of 15 m × 61 m and were georeferenced using the Global Positioning System. These sites were planted early and regularly monitored for disease symptoms every 1 to 2 weeks by state collaborators and extension specialists. Cucurbits grown at the sentinel sites were cucumber cv. Straight 8 and Poinsett 76 (*Cucumis sativus*), cantaloupe cv. Hales Best Jumbo (*Cucumis melo*), acorn squash cv. Table Ace (*Cucurbita pepo*), butternut squash cv. Waltham (*Cucurbita moschata*), giant pumpkin cv. Big Max (*Cucurbita maxima*), and watermelon cv. Micky Lee (*Citrullus lanatus*) (*Ojiambo et al., 2011*). Non-sentinel reports were from locations not designated for regular surveillance but rather voluntary reports from commercial fields, research plots, and home gardens (Table 1). These non-sentinel reports are useful given that, in some epidemic years, CDM was reported earlier in non-sentinel plots than in sentinel plots and thus they could be informative for inferring sources for disease spread.

Latitudes and longitudes geo-coordinates for sentinel and non-sentinel sites were generated from the customized section of the CDM ipmPIPE website (http://cdm.ipmpipe.org). Where no plot data were available, latitudes and longitudes of county centroids were extracted from US Census Bureau 1990 Gazetteer Files and used as approximate georeferenced points. The compiled data from sentinel and non-sentinel sites included, among other things, the date of first disease symptoms, planting type (sentinel plot, commercial field, research plot, home garden, or

unspecified), state, county, and geo-location. A disease case represented a unique combination of host and date of first disease symptoms at a particular location. The total number of disease cases across the study years ranged from 114 to 220, while the number of counties affected ranged from 86 to 179 across epidemic years (Table 1). Correlation analysis was performed to determine whether the number of counties influenced the number of disease reports (Fig. S1) and whether numbers of sites with active surveillance were correlated with the number of counties (Fig. S2) in the region during the study period.

Hourly wind speed and direction at each sentinel plot were derived from weather observations from the National Oceanic and Atmospheric Administration Integrated Surface Database (*Smith et al., 2011*) provided by BASF (Research Triangle Park, Raleigh, NC). Wind measurements were taken at 10 m above the ground. Meteorological wind direction is the direction the wind is blowing from, e.g., wind coming from the north is a northerly wind, and a southerly wind is a wind coming from the south. The raw observations for the meteorological wind direction for a northerly wind is defined as 360°, a southerly wind is 180°, a westerly wind is 270°, and an easterly wind is 90° (Fig. S3). Meteorological wind direction (*wd*) in degrees was converted to a mathematical direction (*md*, i.e., the angle as measured in the mathematically conventional way, counterclockwise from the eastward direction) in degrees using the formula:

$$md = \begin{cases} 270 - wd, & if\ wd \le 270 \\ 360 + (270 - wd), & if\ wd > 270 \end{cases} \quad (1)$$

The mathematical direction in degrees was subsequently converted to radians. The x and y (u and v) components of the hourly wind vectors were then calculated as: $x = r\cos\theta$ and $y = r\sin\theta$, where r is the wind speed in miles per hour and θ is the wind direction in radians (Fig. S3).

Static network analysis

Spatial networks were constructed for each epidemic year to provide insight into the structure of CDM spread in the eastern United States. The general methodology involved linking a 'source' node *i* at a one location to a 'sink' node *j* at another location using a probability based on the distance between the two nodes. This probability is given by a connection kernel which decays with distance such that connections are predominantly localized (*Danon et al., 2010*). In this study, nodes were a combination of sentinel and non-sentinel sites in the eastern United States. We point out that other locations in the eastern United States that were not monitored in this study may contribute to the risk and spread of CDM. However, only the reports of the locations where CDM was monitored or reported were available for inclusion in this study.

Let N be a set of nodes with node $i \in \{1,..., N\}$ and node $j \in \{2,..., N\}$. To form the static network, a link (l) between two nodes (i and j) was determined as a function of distance. Between-node Euclidean distances were calculated using the Haversine formula (*Sinnot, 1984*) using the package geosphere (*Hijmans, 2017*) implemented in the R programming language (R Development Core Team). The x and y displacement vectors for two nodes were calculated based on the equirectangular projection as follows:

$$
\begin{aligned}
z &= \sin^2[(\varphi_j + \varphi_i)/2] + \cos(\varphi_j)\cos(\varphi_j)\sin^2[(\lambda_j - \lambda_i)/2] \\
l_{ij} &= \mathrm{R} \times 2 \times \operatorname{atan2}(\sqrt{z}, \sqrt{1-z}) \\
x &= \mathrm{R} \times (\lambda_j - \lambda_i)\cos[(\varphi_j + \varphi_i)/2] \\
y &= \mathrm{R} \times (\varphi_j - \varphi_i)
\end{aligned}
\qquad (2)
$$

Where φ = latitude (radians), λ = longitude (radians), R = radius of the earth (mean = 6,371 km), and l_{ij} = haversine distance between node i to node j.

Links were created using an inverse power-law dispersal kernel $y = (l_{ij})^{-b}$, where y is the

probability of transmission from node *i* to node *j* (*Andersen et al*., *2019*), l_{ij} is the distance between node *i* to node *j*, and *b* is the spread parameter. The parameter *b* was not estimated in this study but were obtained from a previous study on the isotropic spread of CDM in the eastern United States (*Ojiambo et al*., *2017*) that used the same epidemic data from 2008 to 2016 that was used in the present study. In that study, the authors examined how *b* varied over multiple epidemic years and found that *b* varied over years ranging from 1.61 to 3.36. Thus, values of *b* generated from that study were used in corresponding epidemic years examined in the present study as a representation of isotropic spread through links in the network. In essence, a link was formed between node *i* to node *j* based on whether they are within a certain distance and a link was created between node *i* to node *j* if $y > \tau$ for $0 < \tau < 1$, where τ is the threshold probability of pathogen transmission.

Several static networks were created for a range of values of τ for uncertainty analysis to determine the influence of τ on link formation as described by *Andersen et al.* (*2019*). The range of τ selected was bounded by values that produced a full network and a near-zero probability of link formation (Fig. S4) to facilitate identification of a network with a giant component (GC), since a network without a GC does not provide much information on the behavior of epidemic spread. Thus, the value of τ selected to generate the final static network was identified in two stages. First, τ had to result in a network where each node was connected to at least another node (*Ferrari, Preisser & Fitzpatrick, 2014*). Second, the selected τ also had to have a high proportion of nodes within the GC in the resulting static network. For each epidemic year, the final static network generated using the selected value τ for each epidemic year was used in additional network analyses described below (dynamic networks and error quantification). Degree and the exponent of degree distribution, γ, for final static networks were estimated using R as described by *Kolaczyk & Csárdi* (2020).

Network centrality measures

Centrality measures, betweenness centrality (BWC), closeness centrality (CLC), degree centrality (DGC) and eigenvector centrality (ECV) (Table 2), were calculated using the *igraph* package in R (*Csárdi & Nepusz, 2006*) for each static network that was created for different τ values as described below (identification of important nodes). The empirical cumulative probability distributions of BWC, CLC, DGC, and EVC were calculated for each epidemic year to describe the distribution of the calculated centrality metrics across all nodes. For a set of centrality metrics across a set of nodes, the probability of each value was calculated and the empirical cumulative density function in the *ggpubr* package in R was used to calculate the cumulative probability distributions of BWC, CLC, DGC, and EVC. The similarity in ranking of nodes among centrality metrics was then assessed using Spearman's rank-based correlation.

Identification of important nodes for disease spread within the static network

Analysis of disease outbreaks from 2008 to 2016 was conducted to determine if recurring patterns of disease spread occurred that could help to identify important nodes in the networks. We tallied the number of times a node was observed across epidemic years, i.e., the infection frequency. Two approaches were used to identify nodes potentially important for disease spread and that could be useful for risk-based surveillance or disease mitigation: i) selection of nodes based on infection frequency and ii) selection of nodes based on a combination of infection frequency and centrality metrics.

In the first approach, infection frequency was calculated for nodes in the dataset and nodes were then ranked from highest to lowest based on their infection frequency. In the second approach, the entire dataset was reduced to contain only nodes where disease occurred in at least

one year. A static network was then created such that each node was connected to at least another node (*Ferrari, Preisser & Fitzpatrick*, *2014*) using $b = 2.11$ as estimated previously by *Ojiambo et al*. (*2017*) and the centrality metrics were calculated for this network. Centrality metrics were scaled to a value between 0 and 1 and combined with infection frequency in a ratio of 4:1 (frequency:centrality) for each node to give more weight to infection frequency as described by *Sutrave et al*. (*2012*). Nodes were then ranked in decreasing order based on this weighted value. This weighting approach puts more emphasis on nodes where the disease is observed recurrently and nodes that either are highly connected and acting as bridges to other nodes (BWC), occur on the shortest path (CLC), or connected to other potential super-spreaders (DGC and EVC). A sensitivity analysis was conducted with four additional frequency:centrality ratios with different weights. The results of this analysis showed that changing the weights changed the ranks but did not give more weight to the infection frequency (Fig. S5). Further, of all ratios tested, only the 4:1 ratio resulted in consistent results wherein the higher frequency nodes also had higher weights and were ranked higher (Table S1).

For each epidemic year, a range of threshold values ($0 < \tau < 1$) was considered such that bounds for τ produced a range of dense networks and sparse networks. In each year, 20 individual values of τ were used to construct 20 networks. Centrality metrics were calculated for each network and the results were ranked in a decreasing order. The top 20 nodes with the highest scores were then selected and a second ranking was done for each node in this set. The number of times a node appeared in the top 20 ranking across all thresholds was recorded to eliminate the nodes that were ranked with higher scores in the dense and sparse networks. The nodes were then ranked in decreasing order. The results across centrality metrics and τ values were combined into a heatmap visualization using the *ggplot2* package in R (*Wickham*, *2016*).

Dynamic network model of cucurbit downy mildew

To describe the dynamic process of disease spread occurring on a static network, we modeled the probability of different nodes being infected over a discrete weekly time step, $t \in \{1,2,...,T\}$, in each epidemic year, based on a simplified SI model described by *Sutrave et al*. (*2012*) with the following assumptions: i) the pathogen is primarily dispersed by wind, ii) host response to the pathogen is homogeneous and iii) weather is favorable for infection and disease spread. This model combines the static (constant during each year) and the dynamic (time-varying during each year) components of the network and was formulated as:

$$\begin{cases} \alpha_{ij} = (l_{ij})^{-b} \\ \beta_{ij} = \dfrac{\vec{l}_{ij} \cdot \vec{w}_t}{\left|\vec{l}_{ij}\right|} \\ u_{ij} = \alpha_{ij} \times \beta_{ij} \end{cases} \quad (3)$$

where α_{ij} is a constant function of the between-node distance and decays exponentially with distance, β_{ij} is the dynamic wind-based infection rate, l_{ij} and b are as defined above, $\vec{l}_{ij}$ is the displacement vector between two nodes, $\vec{w}_{ij}$ is the wind vector at time t, and u_{ij} is the link weight based on distance and wind between node i and node j at time t.

Given that the probability of a node being infected depends on the number of infected neighbors, the probability ϑ_i of node i not being infected by its neighbors was calculated as:

$$\vartheta_i(t) = \prod_{j \in N_i} (1 - u_{ij} \cdot p_j(t)) \quad (4)$$

where p_j is the probability of node j being infected at time t, $u_{ij} \in [0,1]$ is the link weight as defined above, and N_i is a set of neighbors of node i. Given Equation 4, the probability p_i of node i being infected at time t was calculated thus:

$$p_i(t) = 1 - (1 - p_i(t-1))\vartheta_i(t) \quad (5)$$

Values of p_i and β_{ij} were calculated and updated, respectively, at each weekly time step. All calculations were performed in MATLAB version R2019a (MathWorks Inc., Natick, MA).

Error quantification in the dynamic network model

The observed infection status of a node and the corresponding predicted infection probability of the node were used to quantify the error in the dynamic network model. First, a value of 0 or 1 was assigned to a node that was either non-infected or infected, respectively, in the observed data at each time step *t*. Secondly, the observed (0 or 1) value for each node was compared to the corresponding infection probability calculated by the model at each time step *t*. The error was then defined as the absolute difference between the observed and predicted infection probability. Mean error for the infected nodes at time step *t* was then calculated as (*Sutrave et al*., *2012*):

$$\hat{E}_{in}(t) = \frac{\sum_{i=1}^{N_{in}(t)} (1 - p_i(t))}{N_{in}(t)} \quad (6)$$

where $N_{in}(t)$ is the total number of infected nodes at time step *t*, while $p_i(t)$ is as defined above. Similarly, the mean error for healthy nodes for at each time step *t* was calculated as:

$$\hat{E}_{hn}(t) = \frac{\sum_{i=1}^{N_{hn}(t)} p_i(t)}{N_{hn}(t)} \quad (7)$$

where $N_{hn}(t)$ is the total number of healthy nodes at time step t. The total error was obtained by using the expression:

$$\hat{E} = \upsilon \hat{E}_{in}(t) + (1 - \upsilon)\hat{E}_{hn}(t) \quad (8)$$

where υ is a weighting factor. The ratio υ: (1- υ) in Equation 8 was 4:1 such that observed-infected nodes were given four times more weight than the observed-healthy nodes in the evaluation of the total error. Here, it was deemed more important to correctly predict infection than to correctly predict an absence of infection such that a few nodes incorrectly predicted will have an insignificant effect on the prediction error (*Sutrave et al., 2012*). All these calculations were performed in MATLAB.

Assessing node importance in disease spread using a dynamic network

The importance of nodes identified as highly connected based on the four centrality measures from the static network analysis, i.e., BWC, CLC, DGC, and EVC, were subsequently evaluated for their impact on disease spread based on link structures of the dynamic network model described above. Nodes identified as most important based on each centrality metric were removed from the networks and the probabilities of disease spread among the remaining nodes were recalculated in the new dynamic network for each epidemic year as described above. Prediction of disease outbreaks based on all nodes present in the network was subsequently compared to prediction of disease outbreaks when nodes identified as important based on the above centrality measures were removed from the network. This approach of node evaluation is equivalent to intensive disease management, where important nodes are completely removed and the resultant impact of their removal on disease propagation within the network is assessed (*Sutrave et al., 2012*). A sensitivity analysis was also conducted for a range of υ:(1- υ) ratios to examine the effect of the choice of the value of υ on the model prediction errors. This analysis showed that increasing the value of υ resulted in negligible changes in prediction errors across all centrality measures and epidemic years (Table S2).

RESULTS

Spatiotemporal dynamics of disease spread in the eastern United States

Observations of disease outbreak suggested a spatial association between the locations of first and last disease reports. The disease was first observed in a sentinel plot in southern Florida in Miami-Dade County in 6 out of 8 epidemic years (Fig. 1). Most of the first disease reports from 2008 to 2016 occurred in February and March in southern Florida or southwestern Texas along the Gulf of Mexico, with reports of initial disease outbreaks being from both sentinel and non-sentinel sites.

Subsequent reports of new disease outbreaks progressed northward with time, with new outbreaks occurring later in more northern states (Fig. 1). The first outbreaks of CDM in more northern states (e.g., Michigan, New York, or Wisconsin) occurred considerably later than corresponding reports of first CDM outbreaks in southern states (e.g., Alabama, Georgia or South Carolina). Across all years, the last set of new disease reports occurred in July, August and September across several states within the region (Fig. 1).

The total number of states with CDM ranged from 22 to 27, and the corresponding number of counties ranged from 86 to 179 across the region (Table 1). There was a positive correlation (r = 0.95; P = 0.0007) between the number of disease reports and counties (Fig. S1), with the number of sites increasing with an increasing number of infected counties. However, the correlation between the number of counties where the disease was present and the number of counties where surveillance was occurring was not significant (r = 0.37; P = 0.3300) (Fig. S2). The linear maximum distance between two disease reports, a measure of epidemic extent, ranged from 2,491 km in 2012 to 3,071 km in 2015.

The number of times that nodes were infected based on combined epidemic data across all years varied from 1 to 6 (Fig. 2). Nodes where the infection frequency was consistently higher

than the median frequency (frequency > 3) were in Alabama, Maryland, Michigan, North Carolina, Ohio, and South Carolina. Nodes with the highest levels of infection frequency were in Wicomico County in Maryland, Johnson, Lenoir, New Hanover, and Sampson counties in North Carolina, and Sandusky, Huron, and Wayne counties in Ohio, with an infection frequency of 5 and 6 (Fig. 2). The remaining nodes had an infection frequency less than the median and they constituted most of the nodes present in counties scattered throughout the region.

Connectivity threshold and static networks of cucurbit downy mildew

The proportion of nodes in the giant component (GC) and the extent of connectedness in a network were used to select the threshold probability of transmission, τ, to generate the final static networks. For example, for the 2008 epidemic data, networks were more connected at $\tau = 6.21 \times 10^{-9}$ (GC = 1.0) than at $\tau = 1.14 \times 10^{-9}$ (GC = 0.92), with other threshold values resulting in either highly or sparsely connected networks. Thus, to achieve a balance in connectivity, $\tau = 6.21 \times 10^{-9}$ was used to generate the final static network for the epidemic data in 2008 (Fig. 3). Similarly, for the 2009 data, networks were more connected at $\tau = 7.83 \times 10^{-9}$ (GC = 0.98) than at $\tau = 1.12 \times 10^{-8}$ (GC = 0.95) with the remaining threshold values resulting in either highly or sparsely connected networks. Thus, $\tau = 7.83 \times 10^{-9}$ was used to generate the final network for disease records in 2009. This logical approach was used to generate the final networks for disease records for the remaining epidemic years from 2010 to 2016. The corresponding values of τ were 1.0×10^{-19}, 4.72×10^{-13}, 2.55×10^{-13}, 1.0×10^{-14}, 2.55×10^{-17}, 1.0×10^{-12} and 1.0×10^{-12}, respectively (Fig. 3). In summary, the threshold for probability of transmission for the final static networks was very low ranging from (1.0×10^{-19} to 7.8×10^{-9}) and the average degree ranged from 12.9 (in 2014) to 52.1 (in 2015). The exponent of the degree distribution (γ) was 2.34 (2008), 1.63 (2009), 2.03 (2010), 1.75

(2011), 1.93 (2013), 1.82 (2014), 2.05 (2015) and 2.14 (2016). Values of $\gamma \geq 2$ indicate that a network is scale-free, i.e., the degrees follow a power-law distribution and the network is characterized by large hubs or nodes with a very high number of links.

Centrality measures and selection of important nodes

Betweenness, closeness, degree, and eigenvector centrality metrics varied between epidemic years. Variability among the 20 most important nodes for each of these metrics was also observed for the final static network constructed within a given epidemic year. Overall, variability among the 20 most important nodes within any epidemic year across the entire study was high for BWC. For example, BWC values ranged from 264.5 to 888.3 in 2008 (Table 3), from 1147.6 to 2415.7 in 2009 (Table 4), and from 237.6 to 1718.2 in 2010 (Table 5). The mean value for the 20 most important nodes as identified by BWC in these respective years was 441.8, 1656.9, and 474, with corresponding standard deviation of 441.1, 896.7 and 1046.9. Variability among the 20 most important nodes as identified based on the other centrality metrics was relatively limited (Tables 3, 4 and 5), with variability among the nodes identified as important based on CLC being the lowest across the entire study period.

The cumulative probability distribution of BWC across the nodes in the examined networks exhibited a power-law distribution. About 85% of the nodes had BWC values < 250, with BWC >1500 being the largest BWC value observed (Fig. S6). In contrast, the cumulative distribution of CLC and DGC was more characteristic of a normal distribution, with the variance of CLC being relatively smaller than that of DGC. The cumulative distribution of EVC followed a Poisson distribution and except for the most important node in each epidemic year (EVC = 1), each other node had an EVC value that was closer to that of one or two other nodes.

Ranking of nodes considered to be important varied among centrality metrics for epidemic years examined (Tables 3 to 5). Spearman's rank-based correlation coefficients were highest between BWC and CLC, with correlations ranging from 0.43 to 0.74 (Fig. S7). Correlations between BWC and DGC or EVC were relatively lower across the epidemic years except between BWC and DGC in 2016, where $r = 0.46$ (Fig. S7). The consistency in the rankings of nodes based on centrality measures was summarized as a heatmap to visualize unique nodes within the networks (Fig. 4). Many nodes overlapped in their rankings among the top 20 important nodes (across all thresholds and centralities) in 2010 (Fig. 4A) and 2014 (Fig. 4C) based on BWC and CLC. However, most nodes overlapped across the four centrality measures in 2011 (Fig. 4B). For example, node 117 in Lewis County, West Virginia, appeared more than 20 times in the top 20 rankings based on BWC and CLC. This same node also appeared more than 10 times in the top 20 ranking of nodes based on DGC and EVC.

Infection frequency and centrality selection of important nodes

Identifying important nodes based on infection frequency and centrality measures of static networks showed some similarities and differences based on the examined centrality metric. The ranking of nodes based on BWC and CLC was generally similar across years, while rankings based on EVC were different from all other centrality measures. Based on BWC, nodes that had a frequency > 4 had the highest calculated values (combined frequency × centrality), with the largest value being 0.82 for the node in Sandusky County in Ohio (Fig. 5), while the lowest weight was 0.13 for a node in Charleston County in South Carolina. Based on CLC, the largest weight for the source was 0.98 for the node in Sandusky County in Ohio that had a frequency > 6, while the node with the lowest weight was that in Miami-Dade County in Florida that had a weight of 0.198.

Similarly, the node in Sandusky County in Ohio had the highest weight of 0.93 based on DGC, followed by nodes in Johnston, Lenoir and New Hanover counties in North Carolina, Wicomico County in Maryland, and Huron and Wayne counties in Ohio that have a frequency of 5 (Fig. 5). Node ranking based on EVC was comparably different from a ranking based on all other centrality measures. A node in Johnston County in North Carolina had the highest weight of 0.84, followed by nodes in Wicomico County in Maryland, Sampson and Johnston counties in North Carolina and Wayne County in Ohio (Fig. 5).

Dynamic network model of disease spread and predicted probability of node infection

The dynamic network model revealed an emerging and evolving network that differed from the static network representation of disease spread (Fig. 6). Generally, similar temporal and spatial patterns were observed in all other years, although the probabilities between nodes in different states and levels of these probabilities differed between years. In all epidemic years, links between nodes closest to the initial disease outbreak (open square) in southern Florida had the highest probabilities of transmission early in the season (i.e., week 10), while the probability of transmission for links between nodes elsewhere in the network was relatively low (Fig. 6). As epidemics progressed in time and space, link probabilities increased for nodes that were more distant from the initial outbreak in more northern latitudes, although probabilities remained relatively low for isolated nodes (Fig. 6).

The probability of infection increased in time and space, with a generally northward expansion of the epidemic front in all years (Fig. 7). Predicted probability of infection increased most during weeks 20 or later. By week 35, the predicted probability increased for most nodes in the eastern United States, with only a relatively few nodes in Illinois and Michigan having a low

infection probability.

Errors in dynamic model and impact of removal of important nodes on model errors

Based on all nodes in the network, mean absolute errors in the dynamic model generated across weekly time steps and averaged monthly from January to August was lowest in 2015 with a value of 0.09 and highest in 2011 with a value of 0.33. The mean absolute error for the dynamic model across the entire study for all the nodes was 0.21 (Table 6).

Removal of nodes identified as important based on BWC, CLC, DGC, and EVC increased the mean absolute errors, indicating the nodes were indeed important for network structure and prediction accuracy. However, the changes in mean absolute errors after node removal varied depending on the specific centrality measure considered. Removal of nodes identified as important by BWC resulted in the largest mean absolute error, 0.32, a 52.4% error rate relative to the base prediction that included all nodes. In contrast, removing nodes identified as important based on CLC, EVC and DGC led to comparatively small increases in mean absolute error (0.24, 0.24 and 0.25, respectively). Thus, model errors due to the removal of nodes identified as important based on BWC were 3 to 4 times higher than errors resulting from the removal of nodes identified as important based on CLC, DGG, or EVC, indicating BWC was superior in identifying important nodes in this data set (Table 6).

The probability of node infection and epidemic progress in the disease network was also affected by the removal of nodes identified as central in the network. Relative to a network with all nodes present, removing nodes identified as important based on BWC reduced the probability of infection of uninfected nodes in the subsequent time step in all epidemic years (Fig. 8). For example, the removal of the nodes in counties in north Florida, Georgia, and South Carolina that

473 were identified as important based on BWC arrested the progression of CDM and infection of
474 nodes in north Florida, South Georgia, and South Carolina in 2009 by week 25 (Fig. 8). We
475 observed a similar pattern of infection probability being meaningfully changed in other years as
476 well when node removal was based on BWC, with the precise change in infection probability
477 varying in specific years. In contrast, removal of nodes identified as central based on CLC, DGC
478 or EVC had a comparably minor impact on the probability of node infection and epidemic progress
479 in all years (Fig. 8).
480

481 **4. Discussion**

482 Estimating the probability and timing of outbreaks in specific sites, and determining where and
483 when the introduction of inoculum can impact the extent of an epidemic, is one of the challenges
484 in predicting the spread of plant diseases and pests (*Meentemeyer et al*., *2011*; *Fitzpatrick et al*.,
485 *2012*). The CDM pathogen can be dispersed over long distances and the disease can spread rapidly
486 under favorable environmental conditions (*Ojiambo & Holmes*, *2011*). In this study, networks
487 were formulated based on historical epidemic records of CDM to establish how connectivity of
488 cucurbit fields influences pathogen dispersal and disease spread in the eastern United States.
489 Multiple low- to high-density static networks were initially generated and analyzed, and networks
490 with biologically-plausible structures and topologies were selected for further analysis. The
491 exponent of the degree distributions for most of the examined networks followed a power-law
492 distribution, indicating that static networks of CDM displayed scale-free properties (*Pastor-*
493 *Satorras & Vespignani*, *2001*), where most nodes had a small number of links, while a smaller
494 number of nodes had a relatively large number of connections. Scale-free connectivity implies the
495 existence of highly connected nodes (hubs) that are responsible for the rapid spread of disease

within the network (*Jeger et al*., *2007*). The transmission probability threshold is low or even absent in scale-free networks (*Shirley & Rushton*, 2005; *Pastor-Satorras & Vespignani*, *2001*) and this may partly explain the low levels of τ observed in the present study. Disease spread in scale-free networks is rapid and models suggest that control of pathogens spreading in such networks should focus on the highly connected sites (*Jeger et al*., *2007*). Thus, targeted sampling of frequently-infected and highly connected sites that are critical in spreading the disease may benefit disease surveillance.

Sites in Florida, Alabama, North, and South Carolina that were infected more frequently in the past may be candidates for disease surveillance. Acquiring the frequency of infection data is a prerequisite, but constant scouting for the disease is expensive. However, once the historical frequency of infection data is available, additional information about network traits is inexpensive to obtain using mathematical models (*Sutrave et al., 2012*). Network centrality metrics such as BWC, CLC, DGC and EVC can facilitate the identification of such highly connected nodes (*Andersen et al*., *2019*; *Gent et al*., *2019*) and aid in evaluating strategies for selecting nodes for surveillance (*Sanatkar et al*., *2015*). Based on a complete static network model, these centrality measures were used to identify highly connected sites for the spread of CDM in the eastern United States. Combining past infection frequency with centrality measures improved the identification of important nodes. For example, DGC, BWC, and CLC produced similar rankings with the infection-based frequency for nodes with an infection frequency greater than four. Although EVC produced a different ranking, nodes with a frequency greater than four still had high weights, thus agreeing with the rankings from the other centrality measures. The combination of frequency-based and DGC was useful in selecting sampling nodes for sentinel plots for soybean rust in the United States (*Sutrave et al., 2012*). DGC is considered the standard measure in network science

519 and is useful for identifying important nodes in static networks of several pathosystems to inform
520 strategic management (*Christley et al., 2005; Gent et al., 2019; Kiss et al., 2006; Xing et al., 2020*).
521 Unlike other centrality measures, DGC is easier to calculate and does not require assessing the
522 entire network (*Christely et al., 2005*). In this study, DGC was ineffective in identifying important
523 nodes compared to BWC. Further, BWC rankings were poorly correlated with those of DGC
524 except for the epidemic data collected in 2016.

525 Betweenness centrality was more useful in identifying the influential nodes in the network
526 as compared to other commonly used metrics. BWC measures the importance of a node by
527 computing how many times a node of interest is on the shortest paths between any two other nodes.
528 This centrality measure has been used to characterize large networks by way of selected nodes
529 since the seminal work by *Granovette* (*1973*). Nodes of high BWC have been used for determining
530 keystone species in food webs, finding clusters and communities, and analyzing the robustness of
531 networks by identifying sensitive points of failure (*Barabási & Bonabeau, 2003; Girvan &*
532 *Newman*, 2002; *Vasas & Jordán*, *2006*). In epidemiology, nodes with high BWC indicates that
533 they are important in disease spread as they act as bridges or 'hubs' to other nodes. Removal of
534 these nodes can contain an epidemic (*Ezeoke et al*., *2018*), as was observed in this study. The
535 observation that BWC was more informative of node importance than other centrality measures
536 emphasizes the need to generate centrality measures that are specific to the disease of interest
537 (*Holme*, *2018*). Invariably, different centrality measures can result in a different ranking profile of
538 important nodes for different pathosystems, possibly due to the inherent differences in the
539 underlying mechanisms of pathogen dispersal and disease spread, landscape connectivity, or other
540 factors (*Dudkina et al. 2023; Holme, 2018; Singer et al., 2022*).

The importance of the highly connected sites in disease spread was further evaluated using a dynamic network model. Mean absolute errors and the probability of infection in nodes across the networks were relatively insensitive to the removal of nodes identified as central by CLC, DGC, and EVC. In contrast, mean absolute errors and the probability of infection in simulated epidemics were quite sensitive to the removal of nodes identified as central based on BWC. This may be related to the physical location of the nodes identified as highly central by the various centrality measures. Removing nodes identified as important based on CLC, DGC and EVC that were located in Pennsylvania, Ohio, and New York did not affect disease progression northward from southern states, whereas removing important nodes in North Carolina largely prevented disease spread. Nodes with high BWC scores were scattered across the region, including in the southern U.S. Removal of these nodes, reduced disease spread, and in some epidemic years, it entirely halted disease spread from most southern states. Most spread of CDM is over relatively short distances of less than 30 km (*Ojiambo & Holmes, 2011*) as the host is planted from south to north. Since BWC is based on the number of shortest paths that pass through a target node, a target node will have a high BWC score if it appears in many shortest paths. Given the relative short dispersal distances of *P. cubensis*, it is plausible that BWC may be better at capturing the dynamics of disease transmission for most of the dispersal events that drive the spread of CDM.

Where resources available for control are limited, targeting nodes with high BWC for treatment has also been found to be an effective strategy in impeding epidemics caused by a disease that spreads rapidly (*Singer et al*., *2022*). The most central nodes identified as important based on BWC were sites in Michigan in the Great Lakes region, Ohio in the Midwest, and Maryland, North Carolina, South Carolina, and Virginia along the mid-Atlantic coast. These states are located along the seasonal transport pathway of *P. cubensis* spores from overwintering locations from the south

(*Aylor, 2003*). Further, most of these have the largest acreage of cucurbit production in the United States. Thus, a combination of spore transport and host density may be a reason for the location of the most central nodes in the above states. These sites could thus be reasonable targets for more intensive sampling for surveillance when collecting reports of new outbreaks within the region. Potentially, more effective disease management in these highly connected sites, such as the strategic deployment of host resistance, could reduce inoculum production that drives infection in neighboring cucurbit fields in the eastern United States.

Unlike the dynamic model used for the spread of soybean rust in the United States (*Sutrave et al*., *2012*), the dynamic model used in this present study incorporated a power-law dispersal gradient characteristic for the long-distance dispersal of plant pathogens. Based on the 2008 and 2009 epidemic data and point-pattern analysis, the dispersal distances for the CDM pathogen were estimated to be up to 390, 737 and 879 km, with 1,000 km being the maximum possible distance of spatial association (*Ojiambo & Holmes*, 2011). Further, *Ojiambo et al*. (*2017*) showed that the spread parameter *b* varied in different epidemics, with the final epidemic extent ranging from 4.16 $\times 10^8$ to 6.44 $\times 10^8$ km^2. Thus, different values of *b* were used in the construction of static networks and in the dynamic network model to account for the difference in spatial spread in each epidemic year. The dynamic network model used in the present study improves on long-distance dispersal by using a flexible threshold for distance to allow for connectivity of nodes that are further apart (*Ferrari, Preisser & Fitzpatrick*, *2014*). However, the model does not account for differences in environmental factors that are likely to influence pathogen dispersal. In addition, accounting for differences in host susceptibility at the different locations could further improve our ability to generalize the findings reported here to different cucurbit host types. Subsequent studies are also needed to establish how unknown disease sources can be imputed in this network modeling

framework and determine how accounting for these unknown sources could influence the network structure and inference made on the location of highly connected sites for disease surveillance reported in this study. Due to the non-random placement of sentinel plots within the monitoring network, these results may not be generalizable and additional studies may be needed to assess how the random placement of sentinel plots could influence the findings reported in this study.

Acknowledgements

The authors wish to thank the collaborators of the Cucurbit Downy Mildew ipmPIPE project for monitoring and reporting disease outbreaks as part of the disease surveillance program.

Data Availability

Publicly available datasets were analyzed in this study and the data underlying the results presented in the study are available from https://cdm.ipmpipe.org/.

REFERENCES

Ames, GM, George DB, Hampson CP, Kanarek AR, McBee CD, Lockwood DR, Achter J, Webb C. 2011. Using network properties to predict disease dynamics on human contact networks. *Proceedings of the Royal Society B* **278:**3544-3550.

Andersen KF, Buddenhagen CE, Rachkara P, Gibson R, Kalule S, Phillips D, Garrett KA. 2019. Modeling epidemics in seed systems and landscapes to guide management strategies: The case of sweet potato in northern Uganda. *Phytopathology* **109:**1519-1532.

Aylor DE. Spread of plant disease on a continental scale: role of aerial dispersal of pathogens. *Ecology* **84:**1989-1997.

610 **Barabási A-L, Bonabeau E. 2003.** Scale-free networks. *Scientific American* **288:**60-69.

611 **Brown JK, Hovmøller MS. 2002.** Aerial dispersal of pathogens on the global and continental
612 scales and its impact on plant disease. *Science* **297:**537-541.

613 **Christley RM, Pinchbeck GL, Bowers RG, Clancy D, French NP, Bennett R, Turner J. 2005.**
614 Infection in social networks: Using network analysis to identify high-risk individuals. *American*
615 *Journal Epidemiology* **162:**1042-1031.

616 **Cohen Y, van den Langenberg KM, Wehner TC, Ojiambo PS, Hausbeck M, Quesada-**
617 **Ocampo LM, Lebeda A, Sierotzki H, Gisi U. 2015.** Resurgence of *Pseudoperonospora*
618 *cubensis*: The causal agent of cucurbit downy mildew. *Phytopathology* **105:**998-1012.

619 **Crowl TA, Crist TO, Parmenter RR, Belovsky G, Lugo AE. 2008.** The spread of invasive
620 species and infectious disease as drivers of ecosystem change. *Frontiers in Ecology and the*
621 *Environment* **6:**238-246.

622 **Csárdi G, Nepusz T. 2006.** The igraph software package for complex network research.
623 *InterJournal, Complex Systems*, 1695. http://igraph.org.

624 **Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, Ross JV, Vernon MC.**
625 **2010.** Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on*
626 *Infectious Diseases* **Volume 2011**, Article ID 284909.

627 **Dudkina E, Bin M, Breen, J, Crisostomi E, Ferraro P, Kirkland S, Marecek J, Murray-Smith**
628 **R, Parisini T, Stone L, Yilmaz S, Shorten R. 2023.** A comparison of centrality measures and
629 their role in controlling the spread in epidemic networks. *International Journal of Control*
630 DOI:10.1080/00207179.2023.2204969

631 **Ezeoke I, Galac MR, Lin Y, Liem AT, Roth PA, Kilianski A, Gibbon HS, Bloch D, Kornblum**
632 **J, Del Ross P, Janies DA, Weiss D. 2018.** Tracking a serial killer: Integrating phylogenetic

relationships, epidemiology, and geography for two invasive meningococcal disease outbreaks. *PLoS ONE* **13:**e0202615.

Ferguson NM, Donnelly CA, Anderson RM. 2001. The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science* **292:**1155-1160.

Ferrari JR, Preisser EL, Fitzpatrick MC. 2014. Modeling the spread of invasive species using dynamic network models. *Biological Invasions* **16:**949-960.

Firester B, Shtienberg D, Blank L. 2018. Modelling the spatiotemporal dynamics of *Phytophthora infestans* at a regional scale. *Plant Pathology* **67:**1552-1561.

Fitzpatrick MC, Preisser EL, Porter A, Elkinton J, Ellison AM. 2012. Modeling range dynamics in heterogeneous landscapes: invasion of the hemlock woolly adelgid in eastern North America. *Ecological Applications* **22:**472-486.

Garrett KA, Alcalá-Briseño RI, Anderson KF, Buddenhagen CE, Choudhury RA, Fulton JC, Hernandez Nopsa JF, Poudel R, Xing Y. 2018. Network analysis: A systems framework to address grand challenges in plant pathology. *Annual Review of Phytopathology* **56:**559-580.

Gent DH, Bhattacharyya S, Ruiz T. 2019. Prediction of spread and regional development of hop powdery mildew: A network analysis. *Phytopathology* **109:**1392-1403.

Girvan M, Newman MEJ. 2002. Community structure in social and biological networks. *Proceedings of the Natural Academy of Sciences of the United States of America* **99:**7821-7826.

Granovetter M. 1973. The strength of weak ties. *American Journal of Sociology* **78:**1360-1380.

Hijmans RJ. 2017. *geosphere: Spherical Trigonometry*. R Package Version 1.5-7. https://cran.r-project.org/web/packages/geosphere/index.html

Holme P. 2018. Objective measures for sentinel surveillance in network epidemiology. *Physical Review E* **98:**022313.

656 **Holme P. 2017.** Three faces of node importance in network epidemiology: Exact results for small
657 graphs. *Physical Review E* **96:**062305.
658 **Holmes GJ, Ojiambo PS, Hausbeck MK, Quesada-Ocampo L, Keinath AP. 2015.** Resurgence
659 of cucurbit downy mildew in the United States: a watershed event for research and extension.
660 *Plant Disease* **99:**428-441.
661 **Jeger MJ, Pautasso M, Holdenrieder O, Shaw MW. 2007.** Modelling disease spread and control
662 in networks: implications for plant sciences. *New Phytologist* **174:**279-297.
663 **Kao RR, Danon L, Green D.M, Kiss IZ. 2006.** Demographic structure and pathogen dynamics
664 on the network of livestock movements in Great Britain. *Proceedings of the Royal Society B*
665 **273:**1999-2007.
666 **Kiss IZ, Green DM, Kao RR. 2006.** The network of sheep movements within Great Britain:
667 Network properties and their implications for infectious disease spread. *Journal of Royal*
668 *Society Interface* **3:**669-677.
669 **Kolaczyk ED, Csárdi C. 2020.** Statistical Analysis of Network Data with R. Second Edition.
670 Springer Nature, Switzerland, AG.
671 **Main CE, Keever T, Holmes GJ, Davis JM. 2001.** Forecasting long-range transport of downy
672 mildew spores and plant disease epidemics. APSnetFeature-2001-0501
673 **May RM, Levin SA, Sugihara G. 2008.** Ecology for bankers. *Nature* **451:**893-895.
674 **Meghanathan N, Lawrence R. 2016.** Centrality analysis of the United States network graph. In:
675 *3rd International Conference on Electrical, Electronics, Engineering Trends, Communication,*
676 *Optimization and Sciences*. 1-6.
677 **Meentemeyer RK, Cunniffe NJ, Cook AR, Joao, JA, Hunter RD, Rizzo DM, Gilligan CA.**
678 **2011.** Epidemiological modeling of invasion in heterogeneous landscapes: Spread of sudden

679 oak death in California (1990-2030). *Ecosphere* **2:**1-24.

680 **Neufeld KN, Keinath AP, Gugino BK, McGrath MT, Sikora EJ, Miller SA, Ivey ML,**
681 **Langston DB, Dutta B, Keever T, Sims A, Ojiambo PS. 2018.** Predicting the risk of cucurbit
682 downy mildew in the eastern United States using an integrated aerobiological model.
683 *International Journal of Biometeorology* **62:**655-668.

684 **Ojiambo PS, Gent DH, Mehra LK, Christie D, Magarey R. 2017.** Focus expansion and stability
685 of the spread parameter estimate of the power law model for dispersal gradients. *PeerJ* **5:**e3465.

686 **Ojiambo PS, Gent DH, Quesada-Ocampo LM, Hausbeck MK, Holmes GJ. 2015.**
687 Epidemiology and population biology of *Pseudoperonospora cubensis*: a model system for
688 management of downy mildews. *Annual Review of Phytopathology* **53:** 223-246.

689 **Ojiambo PS, Holmes GJ. 2011.** Spatiotemporal spread of cucurbit downy mildew in the eastern
690 United States. *Phytopathology* **101:**451-461.

691 **Ojiambo PS, Holmes GJ, Britton W, Keever T, Adams ML, et al. 2011.** Cucurbit downy
692 mildew ipmPIPE: a next generation web-based interactive tool for disease management and
693 extension outreach. Online. *Plant Health Progress* DOI 10.1094/PHP-2011-0411-01-RV.

694 **Ojwang' AME, Ruiz T, Bhattacharyya S, Chatterjee S, Ojiambo PS, Gent DH. 2021. A**
695 **general framework for spatio-temporal modeling of epidemics with multiple epicenters:**
696 **application to an aerially dispersed plant pathogen.** *Frontiers in Applied Mathematics and*
697 *Statistics* **7:**721352.

698 **Pastor-Satorra R, Vespignani A. 2001.** Epidemic spreading in scale-free networks. *Physical*
699 *Review Letters* **86:**3200-3203.

700 **R Core Team. 2018.** *R: A Language and Environment for Statistical Computing.* R Foundation
701 for Statistical Computing, Vienna, Austria.

Sanatkar MR, Scoglio C, Natarajan B, Isard SA, Garrett KA. 2015. History, epidemic evolution, and model burn-in for a network of annual invasion: Soybean rust. *Phytopathology* **105:**947-955.

Singer BJ, Thompson RN, Bonsall MB. 2022. Evaluating strategies for spatial allocation of vaccines based on risk and centrality. *Journal of Royal Society Interface* **19:** 20210709.

Shirley MDF, Rushton SP. 2005. The impacts of network topology on disease spread. *Ecological Complexity* **2:** 287-299.

Sinnot RW. 1984. Virtues of the Haversine. *Sky and Telescope* **68:**159.

Smith A, Lott N, Vose R. 2011. The integrated surface database: recent developments and partnerships. *Bulletin of American Meteorological Society* **92:**704-708.

Sutrave S, Scoglio C, Isard SA, Hutchinson JMS, Garrett KA. 2012. Identifying highly connected counties compensates for resource limitations when evaluating national spread of an invasive pathogen. *PLoS ONE* **7:**e37793.

Thomas A, Carbone I, Choe K, Quesada-Ocampo LM, Ojiambo PS. 2017. Resurgence of cucurbit downy mildew in the United States: insights from comparative genomic analysis of *Pseudoperonospora cubensis*. *Ecology and Evolution* **7:**6231-6246.

Vasas V, Jordán F. 2006. Topological keystone species in ecological interaction networks: Considering link quality and non-trophic effects. *Ecological Modelling* **196:**365-378.

With KA, Gardner RH, Turner MG. 1997. Landscape connectivity and population distributions in heterogeneous environments. *Oikos* **78:**151-169.

Wickham H. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org

Xing Y, Hernandez Nopsa JF, Andersen KF, Andrade-Piedra JL, Beed FD, Blomme G, Carvajal-Yepes M, Coyne DL, Cuellar WJ, Forbes GA, Kreuze JF, Kroschel J, Kumar PL, Legg JP, Parker M, Schulte-Geldermann E, Sharma K, Garrett KA. 2020. Global cropland connectivity: A risk factor for invasion and saturation by emerging pathogens and pests. *BioScience* **70:**744-758.

Table 1 States, number of counties in eastern United States where cucurbit downy mildew was reported, and number of monitoring sites with disease summarized by planting type, during the study period.

Year	Number of states affected	Number of counties	Number of sites by planting type					
			Commercial	Home garden	Research	Sentinel[a]	Unspecified[b]	Total[c]
2008	22	113	68	10	12	59	5	154
2009	24	165	77	26	24	92	1	220
2010	25	118	77	17	24	25	1	144
2011	23	86	57	10	22	28	0	117
2012	25	149	99	20	23	31	0	173
2013	26	179	118	30	23	29	4	204
2014	23	104	53	16	22	20	3	114
2015	27	171	126	15	22	42	4	209
2016	22	107	61	9	19	33	0	122

[a] Sentinel planting type refers to fixed plots, planted early and designated for weekly monitoring.

[b] Unspecified refers to reports where the planting type was not stated when disease was reported in the cucurbit downy mildew monitoring database.

[c] Total number of disease monitoring sites designated as either commercial, home garden, research, sentinel and unspecified plot.

744 **Table 2** Definition of centrality measures in a network model used to study the spread of cucurbit downy mildew in the eastern United
745 States.

Centrality measure	Central node	Relevance in epidemic spread
Betweenness (BWC)	Acts as a bridge to other nodes	Removal of nodes with high betweenness may contain an epidemic
Closeness (CLC)	Lies on the shortest path	Nodes are able to spread disease through a network
Degree (DGC)	Connected to many other nodes	Nodes with high degree may be 'superspreaders'
Eigenvector (ECV)	Connected to other high-degree nodes	Nodes with neighbors having high degree may be 'superspreaders'

746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769

Table 3. Centrality-based ranking of the twenty most important sites in the cucurbit downy mildew network for the epidemic observed in the eastern United States in 2008.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	74	MS	888.3	89	NC	0.0034	131	PA	73	128	PA	1.000
2	118	OH	665.3	118	OH	0.0034	52	MD	72	131	PA	0.994
3	135	SC	608.6	125	PA	0.0034	125	PA	72	134	PA	0.989
4	124	OH	534.1	128	PA	0.0034	128	PA	72	125	PA	0.981
5	39	KY	517.2	130	PA	0.0034	130	PA	72	130	PA	0.974
6	141	TN	507.2	124	OH	0.0034	127	PA	71	99	NY	0.963
7	31	GA	500.4	52	MD	0.0034	134	PA	69	127	PA	0.962
8	89	NC	471.1	134	PA	0.0034	99	NY	66	102	NY	0.953
9	137	SC	470.8	86	NC	0.0033	102	NY	65	96	NY	0.943
10	82	NC	416.6	148	VA	0.0033	96	NY	64	97	NY	0.930
11	91	NC	416.6	150	VA	0.0033	129	PA	64	98	NY	0.926
12	139	TN	375.8	131	PA	0.0033	11	DE	63	100	NY	0.902
13	52	MD	372.1	87	NC	0.0033	97	NY	63	52	MD	0.879
14	75	MS	336.7	88	NC	0.0033	98	NY	63	126	PA	0.858
15	125	PA	324.7	127	PA	0.0033	13	DE	62	129	PA	0.856
16	128	PA	305.4	80	NC	0.0033	100	NY	61	111	OH	0.847
17	136	SC	290.5	78	NC	0.0033	10	DE	59	113	OH	0.847
18	33	GA	290.1	79	NC	0.0033	93	NJ	59	117	OH	0.828
19	29	GA	279.0	151	VA	0.0032	94	NJ	59	120	OH	0.820
20	34	GA	264.5	39	KY	0.0032	133	PA	59	101	NY	0.814
Mean			441.8			0.0033			65.4			0.913
SD			441.1			0.0000			9.9			0.132

[a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC = eigenvector centrality; SD = Standard deviation.

774 **Table 4.** Centrality-based ranking of twenty most important nodes in the cucurbit downy mildew network for the epidemic observed in
775 the eastern United States in 2009.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	34	GA	2415.7	122	NC	0.0012	74	MI	35	109	NC	1.000
2	212	VA	2390.2	132	NC	0.0012	79	MI	35	136	NC	0.979
3	48	KY	2376.2	134	NC	0.0012	82	MI	33	114	NC	0.979
4	154	OH	2152.4	129	NC	0.0012	93	MI	33	118	NC	0.966
5	32	GA	2011.5	124	NC	0.0012	109	NC	33	130	NC	0.960
6	192	TN	1907.7	135	NC	0.0012	158	OH	33	127	NC	0.960
7	186	SC	1803.5	205	VA	0.0012	200	VA	33	211	VA	0.937
8	169	PA	1796.5	212	VA	0.0012	76	MI	32	119	NC	0.913
9	2	AL	1672.3	48	KY	0.0011	90	MI	32	128	NC	0.906
10	180	SC	1605.4	163	OH	0.0011	114	NC	32	207	VA	0.898
11	104	MS	1515.0	164	OH	0.0011	118	NC	32	115	NC	0.891
12	171	PA	1413.6	165	OH	0.0011	136	NC	32	125	NC	0.887
13	103	MS	1351.4	133	NC	0.0011	211	VA	32	126	NC	0.884
14	25	FL	1343.5	192	TN	0.0011	75	MI	31	113	NC	0.882
15	200	VA	1311.5	123	NC	0.0011	83	MI	31	121	NC	0.872
16	153	OH	1259.8	169	PA	0.0011	88	MI	31	120	NC	0.869
17	147	NY	1258.1	171	PA	0.0011	89	MI	31	112	NC	0.869
18	54	KY	1248.4	183	SC	0.0011	91	MI	31	203	VA	0.867
19	101	MS	1158.2	207	VA	0.0011	92	MI	31	200	VA	0.864
20	158	OH	1147.6	203	VA	0.0011	111	NC	31	110	NC	0.850
Mean			1656.9			0.0011			32.2			0.912
SD			896.7			0.0000			2.8			0.106

776 [a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC =
777 eigenvector centrality; SD = Standard deviation.

778 **Table 5.** Centrality-based ranking of twenty most important sites in the cucurbit downy mildew network for the epidemic observed in
779 the eastern United States in 2010.

	Betweenness[a]			Closeness[a]			Degree[a]			Eigenvector[a]		
Rank	ID	State	BWC	ID	State	CLC	ID	State	DGC	ID	State	EVC
1	30	KY	1718.2	30	KY	0.0033	116	OH	56	116	OH	1.000
2	31	KY	1009.3	31	KY	0.0032	103	OH	54	109	OH	0.998
3	65	MS	691.0	116	OH	0.0032	104	OH	54	106	OH	0.997
4	4	AL	577.1	121	PA	0.0032	105	OH	54	110	OH	0.995
5	139	TX	556.0	105	OH	0.0032	106	OH	54	103	OH	0.995
6	77	NC	486.1	103	OH	0.0032	108	OH	54	113	OH	0.995
7	25	GA	469.3	104	OH	0.0032	109	OH	54	114	OH	0.995
8	74	NC	410.1	108	OH	0.0032	110	OH	54	104	OH	0.995
9	13	FL	404.1	110	OH	0.0032	113	OH	54	108	OH	0.995
10	23	GA	342.0	113	OH	0.0032	114	OH	54	61	MI	0.992
11	26	GA	342.0	114	OH	0.0032	61	MI	53	41	MI	0.983
12	5	AL	331.3	107	OH	0.0032	40	MI	52	53	MI	0.983
13	120	PA	305.2	106	OH	0.0031	41	MI	52	60	MI	0.983
14	130	SC	296.8	109	OH	0.0031	48	MI	52	48	MI	0.983
15	138	TX	282.0	120	PA	0.0031	53	MI	52	105	OH	0.977
16	80	NC	264.0	119	PA	0.0031	60	MI	52	42	MI	0.964
17	67	NC	257.1	115	OH	0.0031	107	OH	52	40	MI	0.960
18	117	PA	253.7	61	MI	0.0031	112	OH	52	111	OH	0.960
19	122	PA	246.7	112	OH	0.0031	122	PA	52	112	OH	0.959
20	140	VA	237.6	111	OH	0.0031	42	MI	51	43	MI	0.959
Mean			474.0			0.0032			53.1			0.983
SD			1046.9			0.0000			3.5			0.029

780 [a] ID = node identification number, BWC = betweenness centrality, CLC = closeness centrality, DGC = degree centrality, and EVC =
781 eigenvector centrality; SD = Standard deviation

Table 6 Absolute errors for a network model based on all sites and removal of sites identified as important based on centrality measures used to study the spatio-temporal spread of cucurbit downy mildew in the eastern United States.

		Error after removal of important nodes based on centrality measure[b]			
Year[a]	All nodes	Betweenness	Closeness	Degree	Eigenvector
2008	0.18	0.31	0.22	0.21	0.22
2009	0.27	0.39	0.29	0.28	0.33
2010	0.15	0.23	0.20	0.19	0.20
2011	0.33	0.40	0.35	0.34	0.34
2012	0.27	0.33	0.27	0.27	0.27
2013	0.28	0.45	0.30	0.31	0.31
2014	0.26	0.44	0.36	0.37	0.37
2015	0.09	0.12	0.10	0.10	0.10
2016	0.10	0.17	0.09	0.10	0.10
Mean	**0.21**	**0.32**	**0.24**	**0.24**	**0.25**

[a] For each year, values are means of absolute model errors generated across monthly time steps from January to August.

[b] The 20 most important nodes identified by each centrality measure were removed in the network and the model rerun to calculate the corresponding absolute errors.

Figure 1. Location of cucurbit downy mildew outbreaks in the eastern United States from 2008 to 2016. Locations are color-coded based on the week of the year. Shapes represent the surveillance plot type associated with disease reports during the study period.

Figure 2. Frequency of cucurbit downy mildew outbreaks across all epidemic years from 2008 to 2016 in the eastern United States. Colors represent the frequency (n) of disease cases: red ($n = 6$), yellow ($n = 5$), green ($n = 4$), light blue ($n = 3$), blue ($n = 2$) and pink ($n = 1$). Frequency represents the number of years a node was observed as an infected node (i.e., a location where the disease was reported).

Figure 3. Static networks of cucurbit downy mildew epidemics in eastern United States from 2008 to 2016. Closed circles are nodes where disease was reported (either in a sentinel and non-sentinel plot) and the lines between two nodes are links for the probability of transmission between two nodes calculated based on the power-law dispersal kernel. Thresholds for probability of pathogen transmission ranged from ranged from 1.0×10^{-19} to 7.8×10^{-9} (see text for details). In all years, the initial source of disease outbreak was in Miami-Dade County (open square) in southern Florida.

Figure 4. A heatmap representation of the most important nodes across 20 thresholds for disease transmission across the network and four centrality measures for 2010 (A), 2011 (B) and 2014 (C) networks. Frequency represents the number of times a node appeared in the top 20 ranked list across all evaluated thresholds.

Figure 5. A depiction of node importance based on a combination of frequency of cucurbit downy mildew occurrence in the eastern United States and betweenness, closeness, degree or eigenvector network centrality measures. Frequency represents the number of years a node was observed as an infected node based on epidemic years from 2008 to 2016. Frequency of occurrence and centrality measures are weighted based on a ratio of 4:1.

Figure 6. Evolving network resulting from a dynamic network model for the spread of cucurbit downy mildew in the eastern United States in 2008, 2013, 2014 and 2015. Black circles indicate node centroids of disease outbreak, while the open square is initial source of disease outbreak. Lines are links that have been scaled relative to the probability of transmission by time, with darker and thicker lines indicating higher probabilities of transmission.

Figure 7. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2014 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with a high probability. Blue nodes represent counties predicted to have no outbreak with negligible probability of infection, and all other shades from green to dark red represent increasing probability of disease outbreak. A single node in Texas was reported as infected by Week 10 in the observed data; thus the county was considered infected with probability of one by Week 10.

834 **Figure 8.** Prediction of cucurbit downy mildew outbreaks in the eastern United States by week 25
835 for all nodes present in the network (i.e., prediction) compared to prediction when the 20 most
836 important nodes (based on betweenness, closeness, degree, and eigenvector centrality measures)
837 are removed from the network based on data from epidemics in 2008, 2009, 2013 and 2014.
838 Diamond symbols are nodes identified as important based on each centrality metric. The initial
839 source of disease outbreak is represented by a square symbol in Miami-Dade County in southern
840 Florida.
841
842

Figure 1

Map of location of disease monitoring

Figure 1. Location of cucurbit downy mildew outbreaks in the eastern United States from 2008 to 2016. Locations are color-coded based on the week of the year. Shapes represent the surveillance plot type associated with disease reports during the study period.

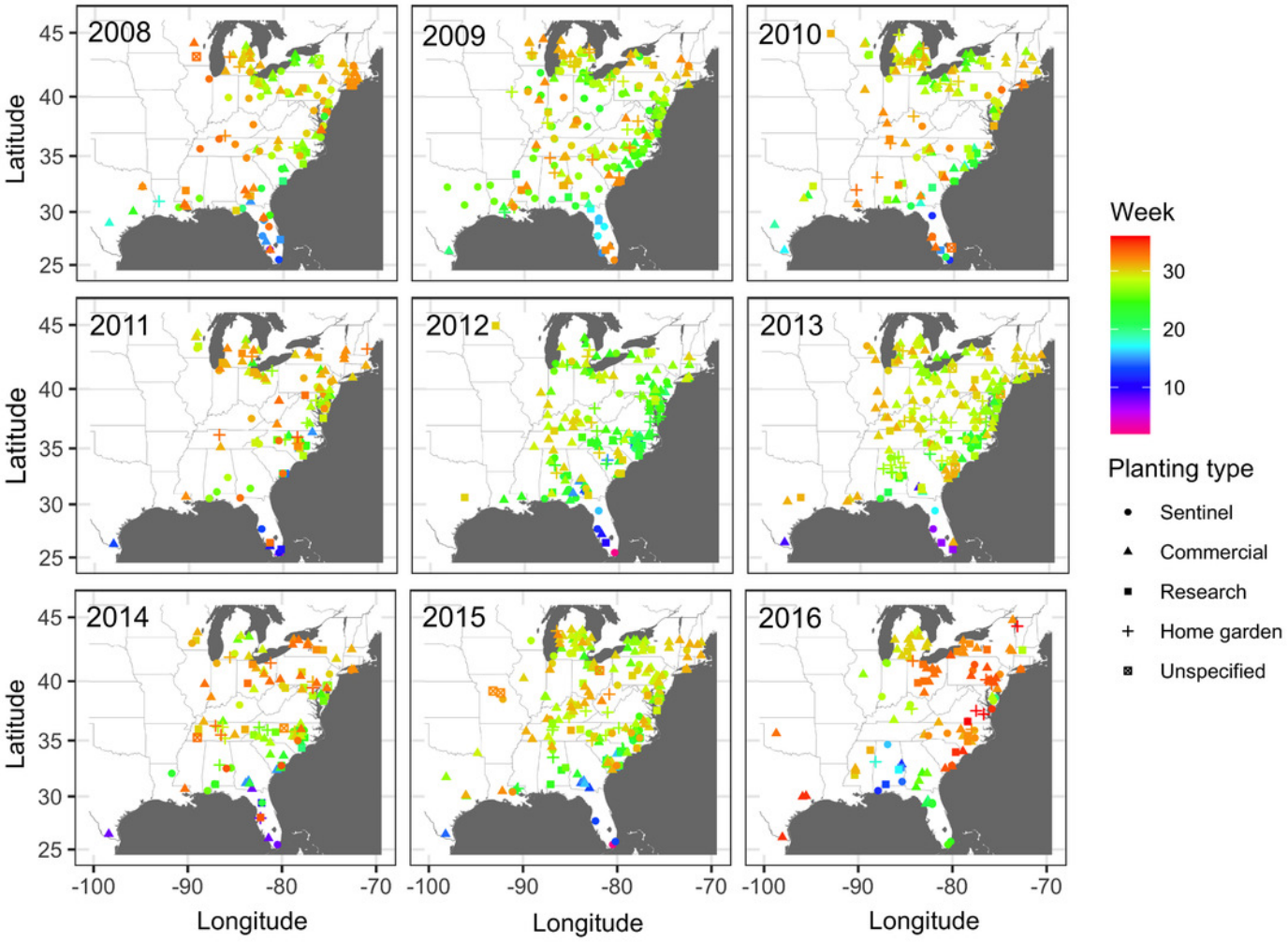

Figure 2

Frequency map of cucurbit downy mildew outbreak

Figure 2. Frequency of cucurbit downy mildew outbreaks across all epidemic years from 2008 to 2016 in the eastern United States. Colors represent the frequency (*n*) of disease cases: red (*n* = 6), yellow (*n* = 5), green (*n* = 4), light blue (*n* = 3), blue (*n* = 2) and pink (*n* = 1). Frequency represents the number of years a node was observed as an infected node (i.e., a location where the disease was reported).

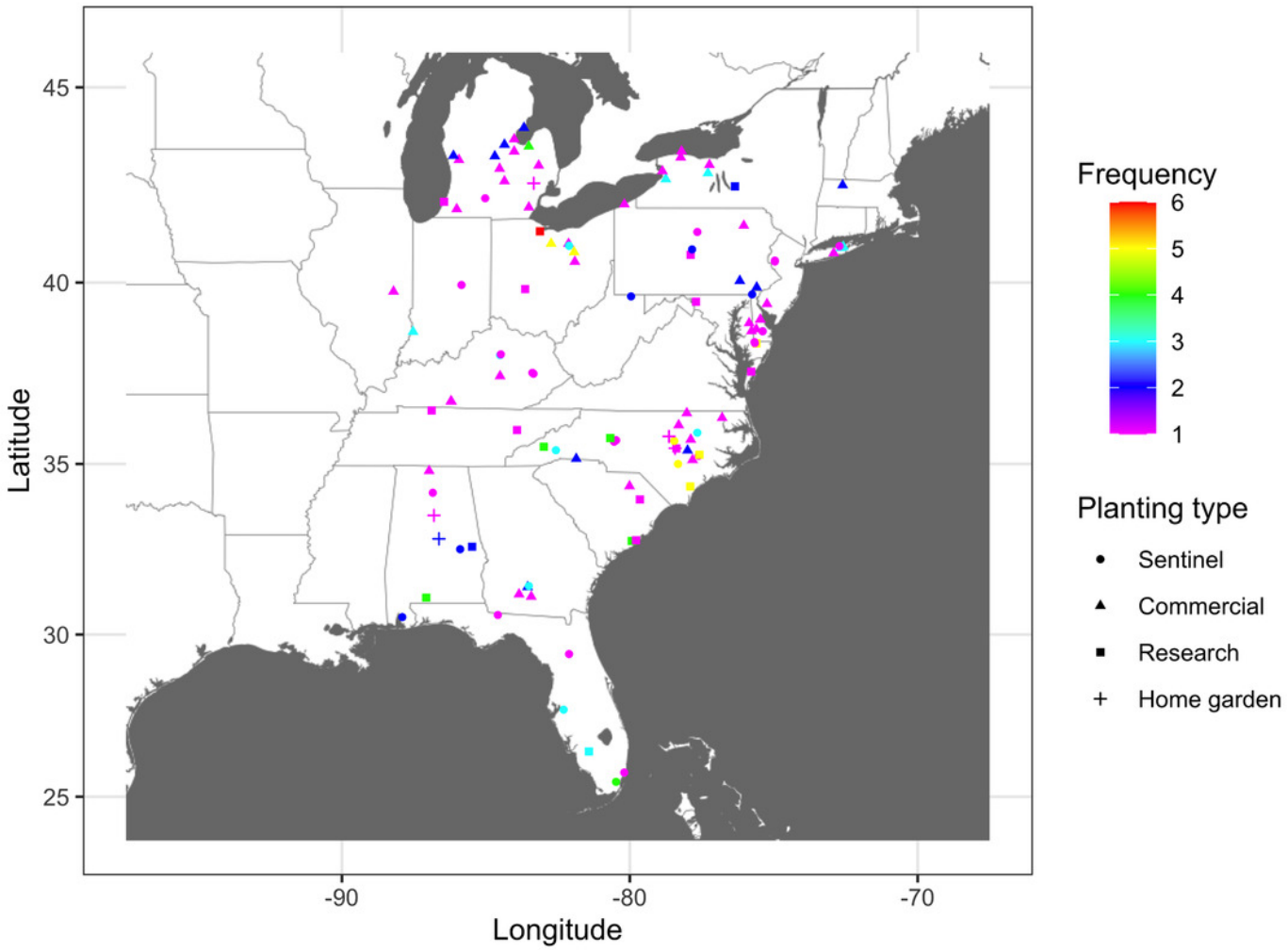

Figure 3

Static networks of cucurbit downy mildew epidemics

Figure 3. Static networks of cucurbit downy mildew epidemics in eastern United States from 2008 to 2016. Closed circles are nodes where disease was reported (either in a sentinel and non-sentinel plot) and the lines between two nodes are links for the probability of transmission between two nodes calculated based on the power-law dispersal kernel.

Thresholds for probability of pathogen transmission ranged from ranged from 1.0×10^{-19} to 7.8×10^{-9} (see text for details). In all years, the initial source of disease outbreak was in Miami-Dade County (open square) in southern Florida.

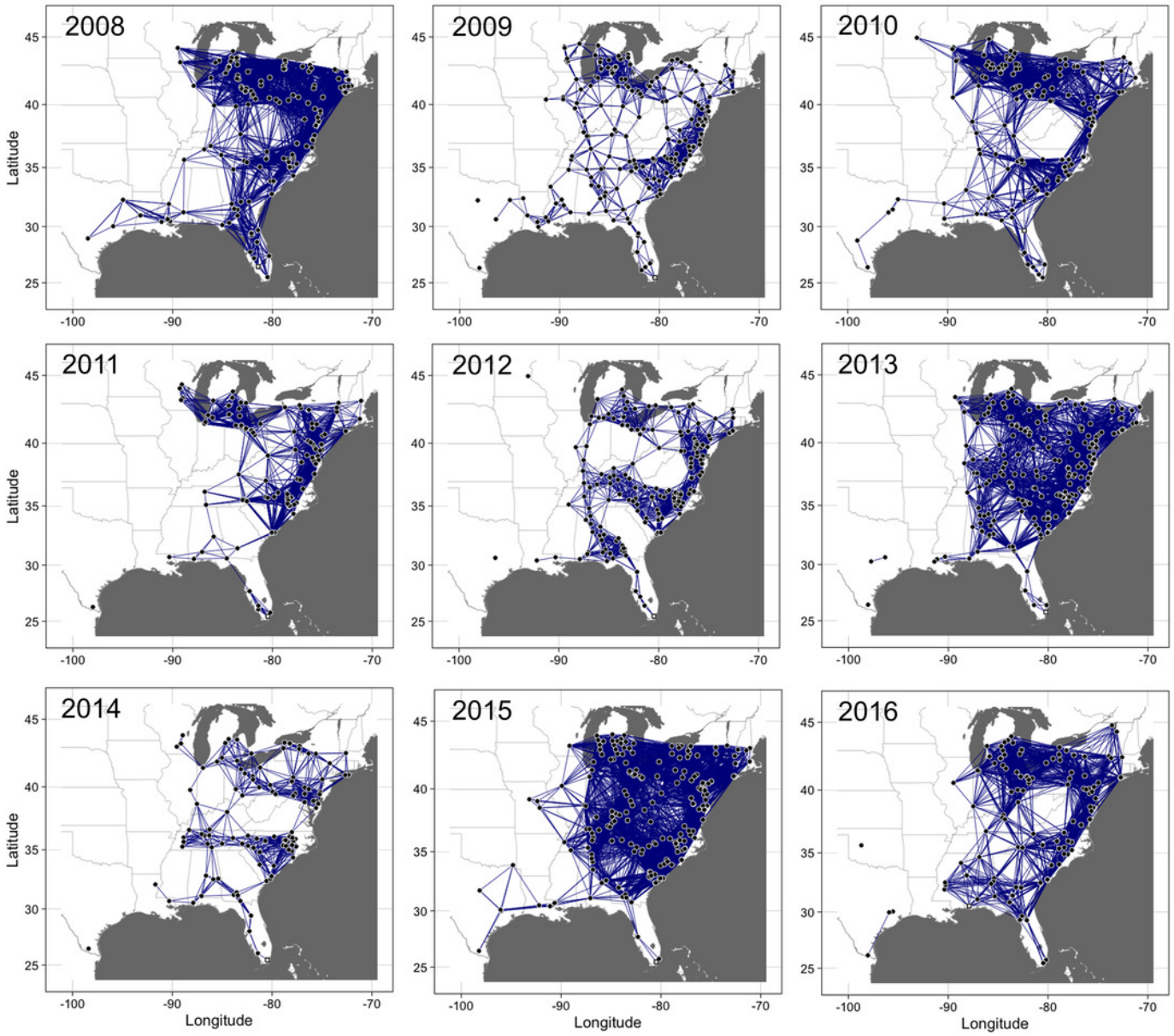
2008
2009
2010
2011
2012
2013
2014
2015
2016
Latitude
Longitude
45
40
35
30
25
-100
-90
-80
-70

Figure 4

Illustration of important nodes for disease spread

Figure 4. A heatmap representation of the most important nodes across 20 thresholds for disease transmission across the network and four centrality measures for 2010 (A), 2011 (B) and 2014 (C) networks. Frequency represents the number of times a node appeared in the top 20 ranked list across all evaluated thresholds.

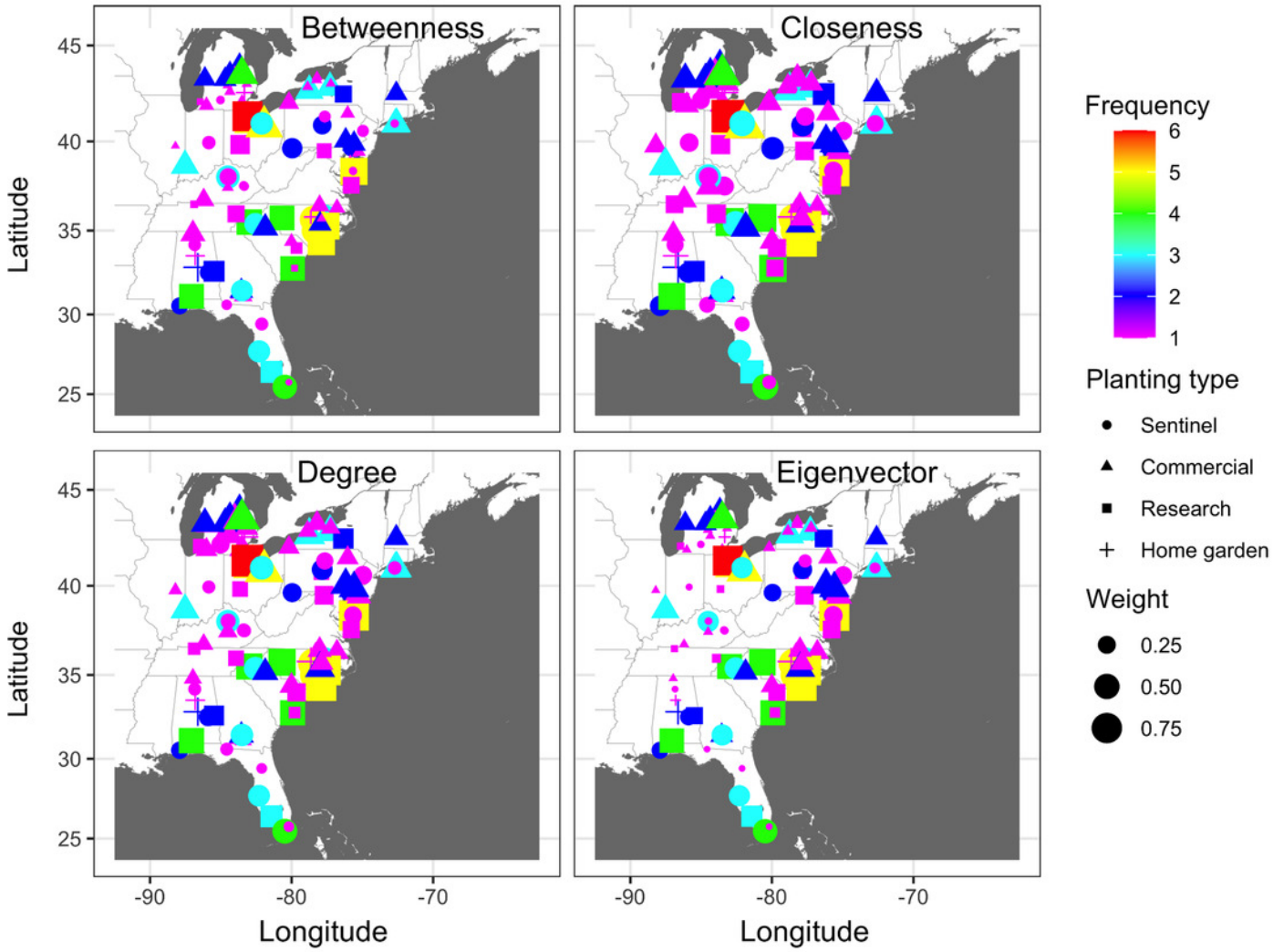

Figure 5

Node importance based on frequency of disease occurence and centrality measures

Figure 5. A depiction of node importance based on a combination of frequency of cucurbit downy mildew occurrence in the eastern United States and betweenness, closeness, degree or eigenvector network centrality measures. Frequency represents the number of years a node was observed as an infected node based on epidemic years from 2008 to 2016. Frequency of occurrence and centrality measures are weighted based on a ratio of 4:1.

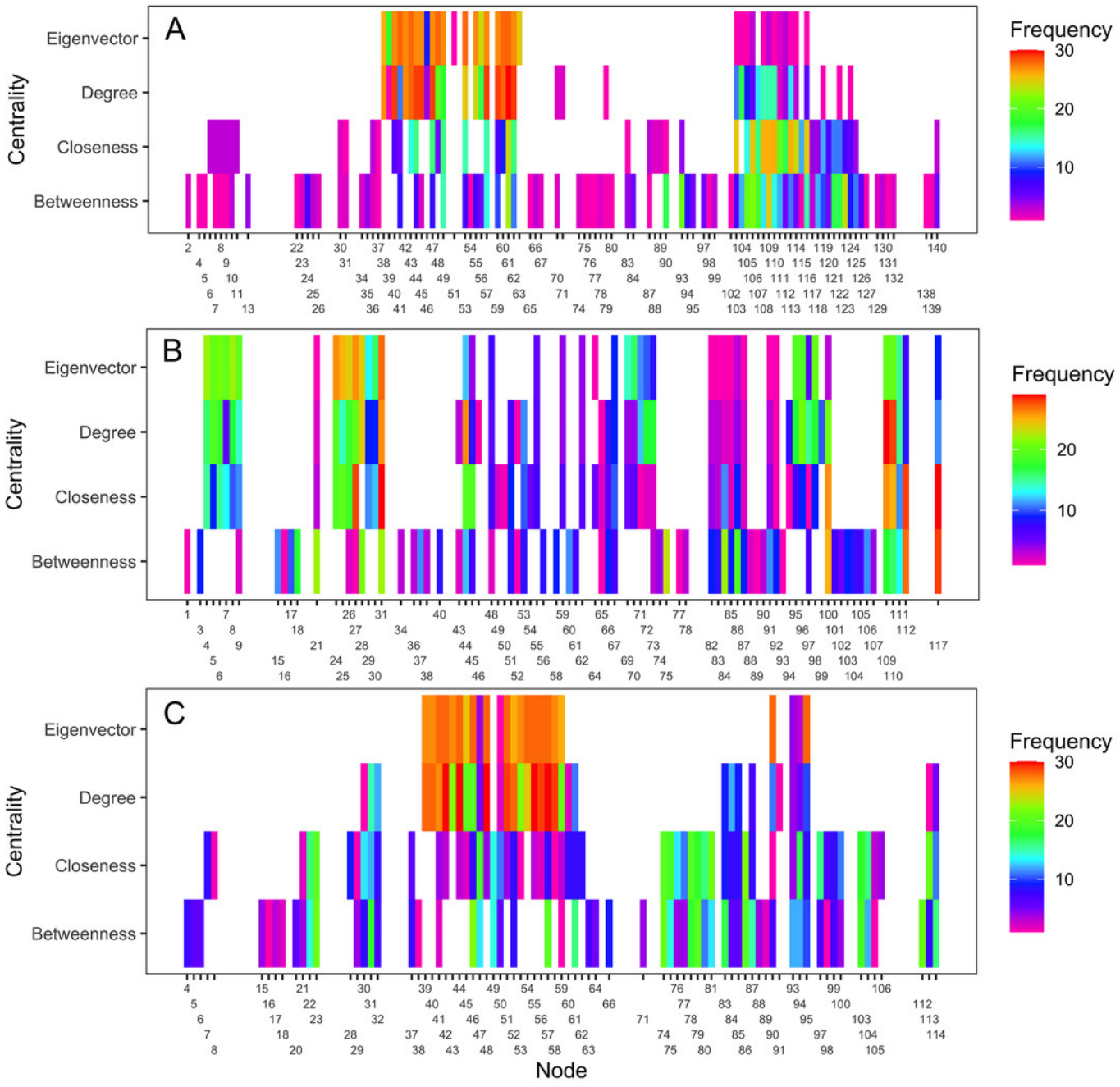
A
B
C
Centrality
Eigenvector
Degree
Closeness
Betweenness
Frequency
30
20
10
Node

Figure 6

Dynamic network of the spread of cucurbit downy mildew

Figure 6. Evolving network resulting from a dynamic network model for the spread of cucurbit downy mildew in the eastern United States in 2008, 2013, 2014 and 2015. Black circles indicate node centroids of disease outbreak, while the open square is initial source of disease outbreak. Lines are links that have been scaled relative to the probability of transmission by time, with darker and thicker lines indicating higher probabilities of transmission.

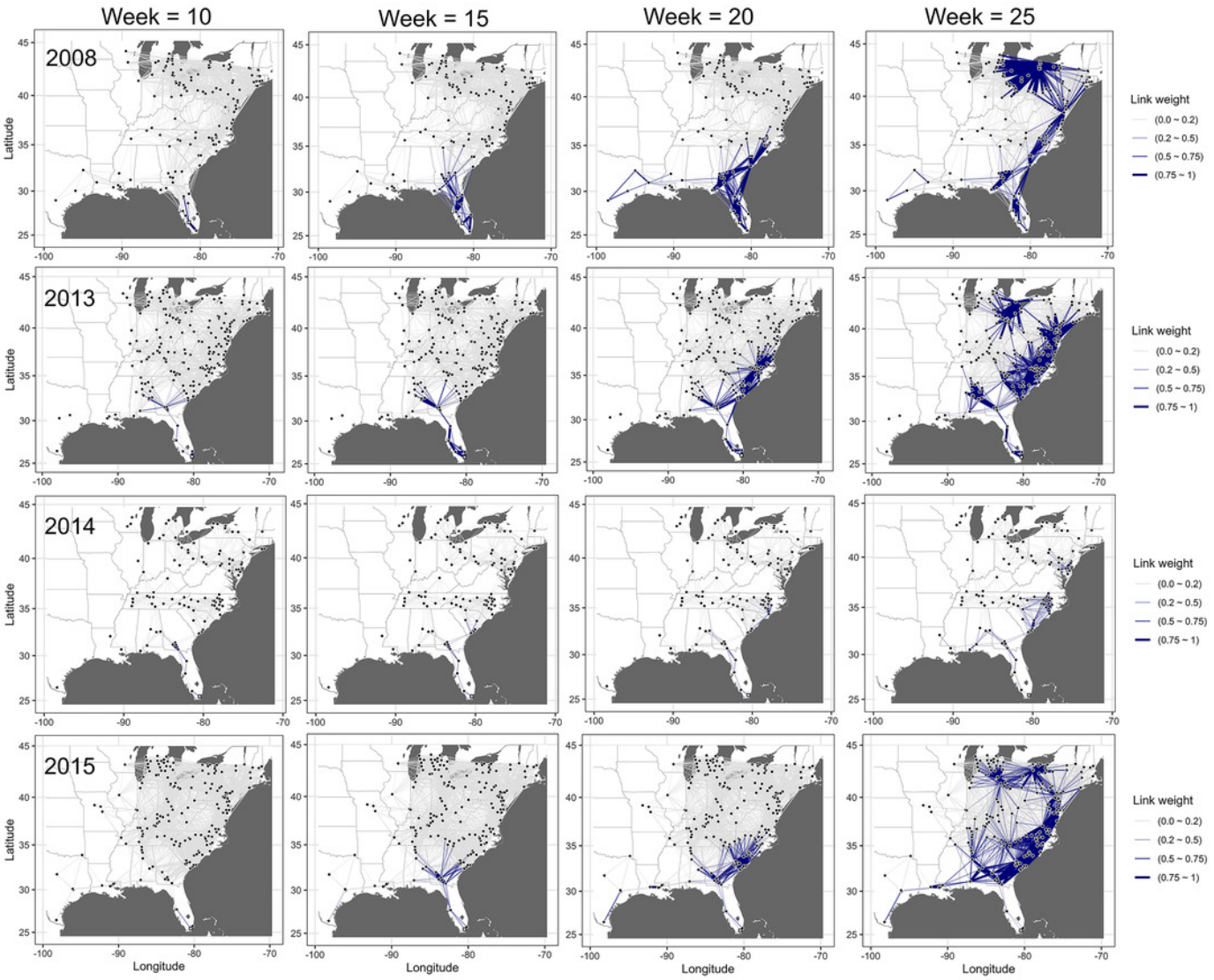
Week = 10
Week = 15
Week = 20
Week = 25
2008
2013
2014
2015
Latitude
Longitude
Link weight
(0.0 ~ 0.2)
(0.2 ~ 0.5)
(0.5 ~ 0.75)
(0.75 ~ 1)

Figure 7

Prediction of the temporal spread of cucurbit downy mildew

Figure 7. Prediction of cucurbit downy mildew outbreaks in the eastern United States in 2014 based on cumulative disease outbreaks observed in previous times steps in the same epidemic year. Dark red nodes represent counties predicted to have an outbreak with a high probability. Blue nodes represent counties predicted to have no outbreak with negligible probability of infection, and all other shades from green to dark red represent increasing probability of disease outbreak. A single node in Texas was reported as infected by Week 10 in the observed data; thus the county was considered infected with probability of one by Week 10.

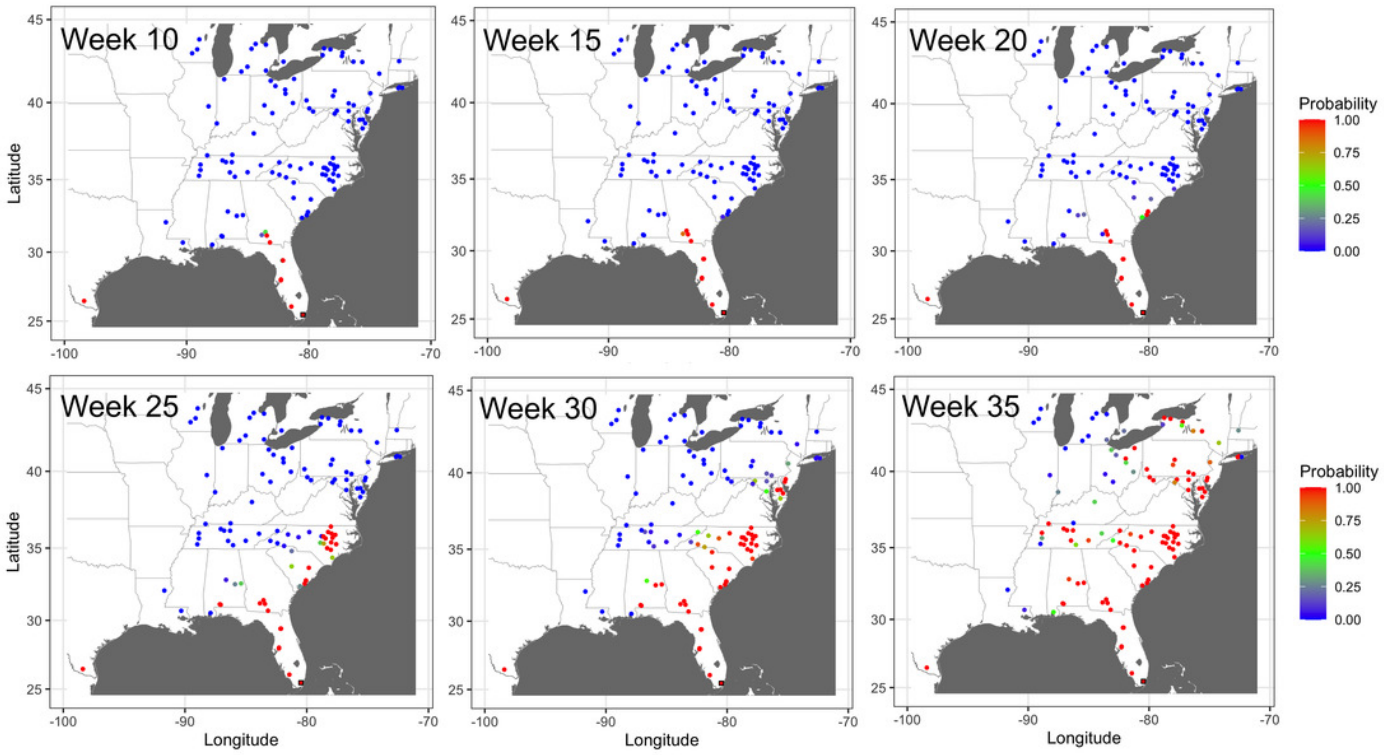

Figure 8

Impact of removal of important nodes on disease spread

Figure 8. Prediction of cucurbit downy mildew outbreaks in the eastern United States by week 25 for all nodes present in the network (i.e., prediction) compared to prediction when the 20 most important nodes (based on betweenness, closeness, degree, and eigenvector centrality measures) are removed from the network based on data from epidemics in 2008, 2009, 2013 and 2014. Diamond symbols are nodes identified as important based on each centrality metric. The initial source of disease outbreak is represented by a square symbol in Miami-Dade County in southern Florida.