Towards Robust Models of Food Flows and Their Role in Invasive Species Spread

Srinivasan Venkatramanan[1,⋆], Sichao Wu[1], Bowen Shi[1], Achla Marathe[1,2], Madhav Marathe[1],
Stephen Eubank[1], Lalit P. Sah[3,4], A. P. Giri[3,4], Luke A. Colavito[3,4],
K. S. Nitin[5], V. Sridhar[5], R. Asokan[5],
Rangaswamy Muniappan[3], G. Norton[2], Abhijin Adiga[1]
[1] *Biocomplexity Institute of Virginia Tech*
[2] *Department of Agricultural and Applied Economics, Virginia Tech*
[3] *Feed the Future Integrated Pest Management Innovation Lab, Virginia Tech*
[4] *International Development Enterprises, Nepal*
[5] *Indian Institute of Horticultural Research*
[⋆] *email:vsriniv@vt.edu*

***Abstract*—We develop a general data-driven methodology that yields network representations of agricultural flows pertaining to the spread of invasive species. The methodology synthesizes sparse, diverse, noisy and incomplete data that is typically available to build realistic spatio-temporal network representations. We illustrate the methodology by modeling the seasonal flow of the tomato crop in Nepal between major domestic markets. Through dynamical analysis of the network, we study its role in the spread of a major pest of tomato, *Tuta absoluta*, an emerging outbreak in this country. In the absence of high-resolution pest distribution data, we apply a novel ranking-based inference approach to establish that tomato trade is a driving factor in the rapid spread of this pest.**

I. Introduction

Food security is an increasingly important societal problem. Increased globalization, climate change, population growth, scarce per capita resources, international trade, travel and invasive species are important factors contributing to the issue of global food security. In this paper, we will focus on commodity flows, a quintessential component of our food systems. Production and consumption of agricultural produce is no longer a local phenomenon – agro products travel thousands of miles over global supply chain networks. While economically attractive in the short term, global trade increases the risk of rapid spread of invasive species and bio-terrorism. The situation is quite similar to spread of infectious diseases in human and animal populations.

In this paper, we study the seasonal flow of agricultural commodities, focusing on their role in the spread of invasive species. The spread of pests and pathogens is driven by various natural and anthropogenic factors. An in-depth understanding of the biology and climatic conditions is essential to assess establishment risk and devise sustainable management strategies and has been the focus of ecologists for a long time. In contrast, not much is understood as regards to the role of human-mediated pathways (including trade and travel) in preventing introduction and mitigating immediate impact [2, 4, 6, 10]. See [7, 8, 13, 17] for further discussion on this important subject.

Our contributions: Here we develop an integrated methodology that combines data science, algorithmics, machine learning and ecological modeling that allow us to address important factors that affect the human mediated pathways contributing to invasive species spread. Our key contributions in this paper are as follows:

(*i*) We develop an integrated data-driven methodology for synthesizing realistic spatio-temporal networks of seasonal agro-products between major markets. The methodology is outlined in Figure 1a. It combines diverse multi-type, noisy, misaligned and sparse datasets with detailed context specific domain knowledge provided by local experts. A particular challenge we address is data sparsity. The methodology is generic and can be adapted to a other agro products and regions.

(*ii*) We illustrate the methodology by developing a spatio-temporal domestic tomato trade network in Nepal and investigate its role in the spread of *Tuta absoluta*, a devastating pest of the tomato crop [3] and an emerging pest in Nepal [1].

(*iii*) We analyze the spatio-temporal properties of the flow networks. Further, through dynamical analysis of the networks and a novel rank-based inference approach, we assess the role of trade in the spread of the pest.

(*iv*) We conduct an in-depth sensitivity analysis to quantify the role of input parameters. This analysis is used in validating our synthesized networks; furthermore the analysis provides improved understanding of the pest dynamics.

Challenges. Agro-trade networks for moving agricultural products is a complex system. The networks depend on varied factors, including seasonal production, population distribution, cultural factors, economic activity, storage and transport infrastructure. Furthermore, data needed to develop agro-trade networks is often sparse, noisy and is not openly available. For instance, even standard information such as region-level production is unavailable for many countries. Even if available, these datasets vary in format, they are misaligned in reporting time and vary in spatial and temporal resolution. Apart from quantitative datasets, there is also need for qualitative information pertaining to the study region such as cultural practices, seasonal production cycles, etc. Interpreting this data and integrating it into model design requires local knowledge. Similar challenges exist in obtaining high-resolution pest distribution data.

Another challenge is validating the network representations. While international trade data is available at the commodity level, domestic data is hard to come by. Even in data-rich regions such as the US, the available sample data (e.g. Freight Analysis Framework[1]) is aggregated at the commodity category level. Secondly, the role of these networks in the study of invasive species requires one to understand the ecological contagion processes. Also, monitoring is a resource intensive task: the placement of traps is largely determined by accessibility and availability of trained personnel. The pest might not be detected during off season due to host unavailability. In the absence of monitoring, its presence will become apparent only during the growing season. But its reporting might be delayed by farmers due to lack of awareness or fear of quarantining. Given these constraints, there may be several months of delay in reporting.

Related work: In recent years, there has been a lot of interest in studying the role of international trade and travel in invasive species spread. Ercsey-Ravasz et al. [8] analyze the International Agro-Food Trade Network to identify countries of importance in the context of food safety. Early et al. [7] study the terrestrial threat from invasive species and evaluate national capacities to prevent and manage invasions. Tatem [17] showed that the world-wide airline network increases the risks of establishment by providing busy transport links between spatially distant, but climatically similar regions of the world.

There has been some work on domestic commodity flow and its role in pest spread. Nopsa et al. [13] evaluated the structure of rail networks in the US and Australia for pest and mycotoxin dispersal. Colunga-Garcia et al. [5] use the regional freight transport information to characterize risk of urban and periurban areas to exotic forest insect pests in the US. In [15] provides a survey of recent modeling efforts.

T. absoluta. There is general consensus that vegetable and seedling trade is a primary driver of *T. absoluta* spread [3]. However, previous modeling efforts have only focused on establishment potential [18] and spatial dispersion [9]. This is the first work that analyzes human-mediated pathways in the context of *T. absoluta*. Nepal's vegetable production and trade has been extensively studied from a socio-economic perspective ([19] for example), but, to the best of our knowledge, there is no such work in the context of invasive species spread with focus on this region.

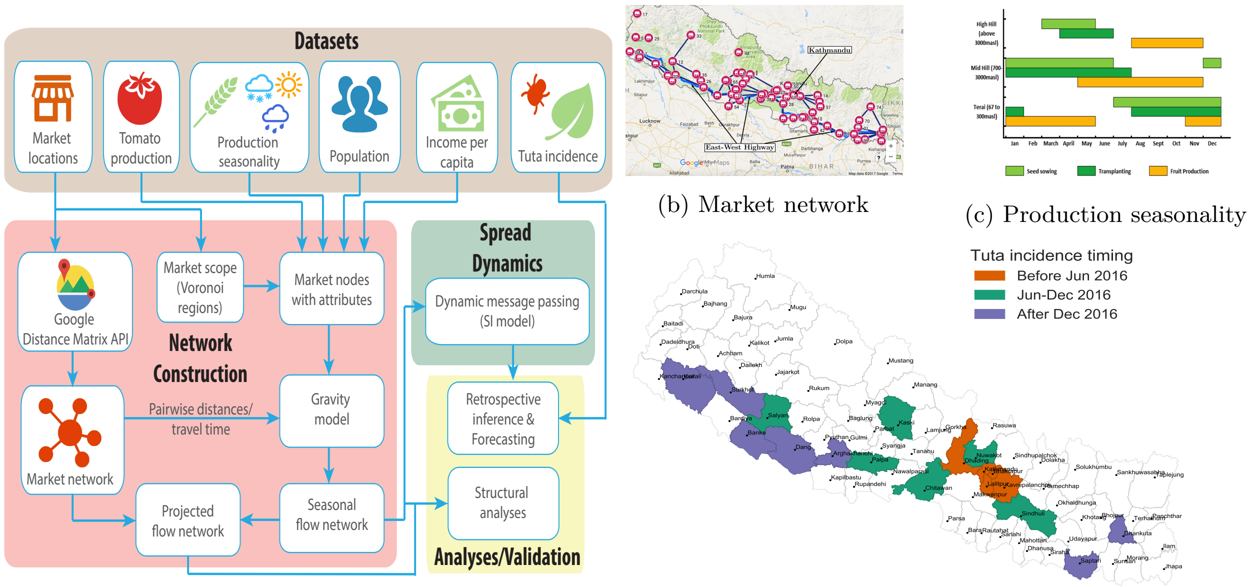
II. Modeling Framework

Figure 1a outlines the different components that constitute the framework. As we describe each component, we will also discuss the associated data challenges and key modeling assumptions that allowed us to integrate them. The symbols and abbreviations used henceforth are summarized in Table II.

A. Data

Table I lists the datasets used in our framework. We link several open-source datasets along with qualitative inputs from local experts in order to model the seasonal trade of tomato crop as well as pest dynamics. Some of the challenges arise from the fact that the datasets vary in their spatial and temporal resolution and their year of release (see Table I). Owing to the unique geography of Nepal, the vegetable production cycle varies with altitude (see Figure 1c). The annual production data was combined with the knowledge of production cycle to model the spatio-temporal variation in

[1] https://ops.fhwa.dot.gov/freight/freight_analysis/faf/



(a) Modeling framework

(b) Market network

(c) Production seasonality

(d) *T. absoluta* incidence timings

Figure 1: Modeling framework and some datasets

production across seasons. Major vegetable markets were geolocated using Google Maps, and Google Distance Matrix API was used to construct the road network and compute travel times. Several organizations have been involved in the monitoring of *T. absoluta* spread in Nepal: NARC, USAID, iDE Nepal, ENBAITA and Agricare Pvt. Ltd. The pest is monitored using pheromone traps that have been installed in several Village Development Committees. In May 2016, *T. absoluta* was officially reported by NARC's entomology division in Lalitpur (near Kathmandu).

B. *Network construction*

Regional markets serve as key locations facilitating agricultural commodity flow, hence it makes sense to model the flow network with markets represented as nodes. We model the flow of agricultural produce among markets based on the following premise: The total outflow from a market depends on the amount of produce in its surrounding regions, and the total inflow is a function of the population it caters to and the corresponding per capita income. The main assumptions in this model are: (i) imports and exports are not significant enough to influence domestic trade: For instance, in 2014, Nepal exported only 1% of its tomatoes and imported about 6-7% of its total consumption (http://www.fao.org/faostat); (ii) Fresh tomatoes are mainly traded for consumption: The tomato processing industry in Nepal is not well developed [19]. This motivates the use of population and per capita income as indicators of tomato consumption in a given district. and (iii) the higher the per capita income, the greater the consumption: Tomato is among the top two vegetables which provide highest profit to farmers (expensive for a typical consumer), and unlike cauliflower and cabbage, tomato is not considered a staple vegetable in the Nepalese household [19].

The flows are estimated using a doubly constrained gravity model [11]. The flow F_{ij} from location i to j is given by

$$F_{ij} = a_i b_j O_i I_j f(d_{ij}) \qquad (1)$$

where, O_i is the total outflow of the commodity from i, I_j is the total inflow to j, d_{ij} is the time taken to travel from i to j, $f(\cdot)$ is the *distance deterrence function*, and coefficients a_i and b_j are computed through an iterative process to ensure flow balance.

However, as seen in Table I, data pertaining to these quantities are available at different spatial and temporal resolutions. Thus, before we apply (1), we need to synthesize these datasets to capture the seasonal commodity flow at the level of markets. The steps involved are described as follows:

Seasonality of production: Based on the physiography, districts of Nepal are partitioned into three regions, namely Terai, Mid Hills and High Hills

Table I: **Datasets.**

Description	Source	Resolution	Year
Population	Nepal Central Bureau of Statistics (http://cbs.gov.np/)	District/Town	2011
Per Capita Income	Nepal Central Bureau of Statistics (http://cbs.gov.np/)	District	2011
Tomato production	Nepal Ministry of Agricultural Development (MOAD) (http://moad.gov.np/)	District, Annual	2015
Production seasonality	iDE Nepal (http://idenepal.org/) and MOAD	Region, Monthly	2016
Major vegetable markets	MOAD Marketing Information System (http://www.agrimis.gov.np/)	Town	2017
Market distances	Google Maps, Distance matrix API	Market	2017
Tomato import/exports	Food and Agriculture Organization (FAOSTAT) (www.fao.org/faostat/)	Country, Annual	2013
Tomato consumption	FAOSTAT, MOAD	Country, Annual	2013
Flows to Kalimati market	Official website (kalimatimarket.gov.np/)	District, Annual	2015
T. absoluta incidence reports	Nepal National Agriculture Research Council(http://narc.gov.np/) USAID IPM Innovation Lab, iDE Nepal	District/town	2017

Table II: **Notation and abbreviations.**

Variables	Description
F_{ij}	Commodity flow from node i to j
O_i	Total outflow of commodity from node i
I_i	Total inflow of commodity into node i
d_{ij}	Distance between nodes i and j
$f(.)$	deterrence function
β	Power-law exponent of gravity model
κ	Cutoff time of gravity model
γ	Per capita income parameter
σ	Gaussian parameter for spatial seeding
t	Time step for the spread model

(see Figure 3e) Due to altitude and temperature variations, the tomato production season varies among these regions (see Figure 1c). Production in the Mid Hills and High Hills is largely restricted to the summer months of June to November (referred to as season S1), while Terai region produces during the winter months of December to May (referred to as season S2). As a result, we have two distinct flow networks, one for each season. We partitioned the districts into two groups: Mid Hills and High Hills belong to group 1, while the Terai districts belong to group 2. All districts belonging to group i were assigned their respective annual production for season Si and zero for the other season.

Market scope definition: The nodes of the flow network are the major markets, 69 in all, after merging markets that belong to the same town. Recall that the amount of production is specified at the district level. In order to obtain the production estimates at market level, we defined *market scope* as follows: The country's map was overlaid by a grid cell of size $5km \times 5km$ and we constructed a Voronoi partition of these cells using market locations as centroids. This is under the assumption that tomato sellers and buyers will seek out the nearest market. We assumed uniform spatial distribution of production within each district. Each grid cell was assigned a value of production in a particular season proportional to the fraction of the area of the district covered by the cell. The total outflow from the market is the sum of production of the grid cells assigned to it for a particular season.

Modeling consumption: We modeled the total inflow I_i into a market as a product of the population catered to by the market and a function of the average per capita income associated with the market η_i, η_i^{γ}, where γ is a tunable parameter. The population catered to by the market, was derived from district level population data and the market scope as defined for production redistribution.

Inter-market travel time: Owing to the diverse landscape of Nepal and varying road conditions we used travel time by road instead of the geodesic or road distance between the markets. We begin with list of major vegetable markets in Nepal (see Table I), and geolocate them using Google Maps. We then manually embedded the market locations onto Nepal road network, and constructed a planar network by connecting the markets which have a direct route (without going through other markets) between them. We also removed markets which were completely inaccessible by road. We used Google Distance Matrix API[2] to compute travel times by road along the edges of this planar network. This in turn, yields a road network among the markets, where the edges are weighted by their travel time. Distance between any two markets is then obtained as the shortest travel time on the road network. The distance deterrence function $f(d_{ij}) = d_{ij}^{-\beta} \exp(-d_{ij}/\kappa)$ combines power-law and exponential decay with d_{ij} which can be controlled by the tunable parameters β, the power-law exponent, and κ, the cutoff time.

C. Spread Dynamics

We develop a discrete-time SI (Susceptible-Infected) epidemic model on directed weighted networks [14] to model pest dispersal. Each node is

[2] https://developers.google.com/maps/documentation/distance-matrix/

either susceptible (free from pest) or infected (pest is present). Henceforth, we use the term "infected" for a node or a region frequently to imply *T. absoluta* infestation at that location. A node i in state I infects each of its out-neighbors j in the network with probability proportional to the flow F_{ij} at each time step t. The infection probabilities are obtained by normalizing flows globally: $\lambda_{ij} = \frac{F_{ij}}{\max_{i,j} F_{ij}}$. The model is based on two assumptions: (i) an infected node remains infected and continues to infect its neighbors and (ii) the chance of infection is directly proportional to the volume traded. Considering the fact that Nepal was ill-prepared for this invasion and the lack of effective intervention methods, (i) is a fair assumption. Historically, *T. absoluta* has spread rapidly in regions where tomato trade has been the highest (parts of Europe and Middle-East for example) thus motivating assumption (ii).

Let $P_S(i, t, f_0)$ denote the probability that node i remains uninfected (i.e., susceptible) by time t given the initial condition f_0 which assigns probability of infection at time step $t = 0$ to each node. In general, computing P_S is hard. Efficient methods have been proposed to estimate this probability. Here, we adopt the *dynamic message passing algorithm* by Lokhov et al. [12], summarized by the following equations.

$$\begin{aligned} P_S^{i\to j}(t+1) =& P_S(i,0,f_0)\Pi_{k\in\delta i\setminus j}\theta^{k\to i}(t+1) \\ \theta^{k\to i}(t+1) =& \theta^{k\to i}(t) - \lambda_{ki}\phi^{k\to i}(t) \\ \phi^{k\to i}(t) =& (1-\lambda_{ki})\phi^{k\to i}(t-1) \\ & - [P_S^{k\to i}(t) - P_S^{k\to i}(t-1)] \end{aligned} \tag{2}$$

In the above equations, λ_{ki} is the infection probability across edge (k, i), and θ, ϕ are intermediate messages used to update the node states. Finally, the quantity of interest $P_S(i, t, f_0)$, the probability that node i remains uninfected (i.e., susceptible) till time t is given as:

$$P_S(i, t+1, f_0) = P_S(i, 0, f_0)\Pi_{k\in\delta i}\theta^{k\to i}(t+1)$$

Note that for any given t, $P_S(i, t, f_0)+P_I(i, t, f_0) = 1$, and hence the entire evolution of the epidemic on the network is captured by $P_S(i, t, f_0), \forall i, t$ given the initial condition f_0.

The initial configuration f_0 is chosen to mimic a spatially dispersed seeding scenario. We first select a *central* seed node, and then use a Gaussian kernel with parameter σ around the seed node to assign initial infection probabilities for neighboring markets. A market at a geodesic distance d from the seed, is assigned the infection probability $e^{-\frac{d^2}{2\sigma^2}}$. The kernel accounts for factors such as uncertainty in determining the pest location, the possibility of spread of the pest through natural means as well as interactions between these markets.

III. Analyses and Results

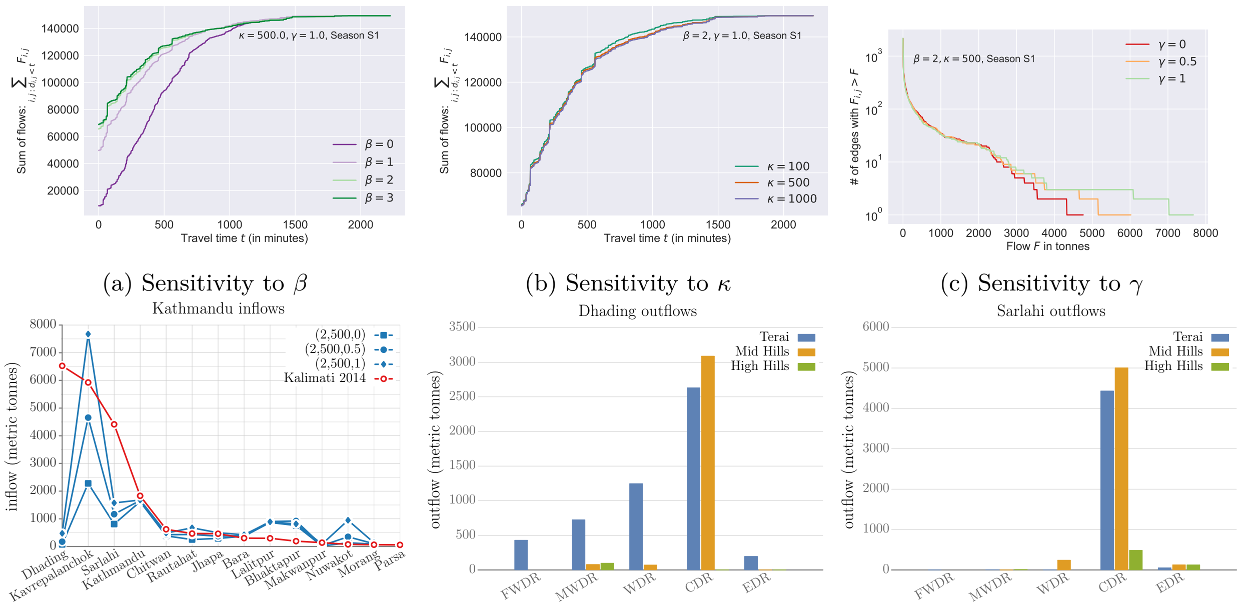
Flow validation: The unavailability of sample data on seasonal trade of tomato crop makes it challenging to calibrate and validate the flow network model. In fact, to the best of our knowledge, even information on annual flow of vegetables between markets is not available. However, for the largest wholesale market of Nepal, Kalimati (located in Kathmandu), yearly data on volume of tomato arriving from each district is available (Table I). In Figures 2d–2f, we compare this data with the network flows. Given a set of network parameters (β, κ, γ), we obtained the inflow from a particular district to Kathmandu as follows: We combined the weights of all edges of the corresponding network with destination node "Kathmandu" and source nodes belonging to that district.

As seen in Figure 2d, for γ values between 0.5 and 1, the flows from the networks are comparable to the Kalimati data except for two districts: Dhading (the top contributor) and Sarlahi (third highest). Upon further investigation we find that Dhading, which is a major producer west of Kathmandu, serves the Mid Hills and Terai regions of the Central Development Region in the flow networks (Figure 2e). While the gravity model predicts that these flows will be directly delivered to these regions, in reality, it is possible that Dhading's produce is routed through Kalimati market as there are several traders from Dhading registered in the Kalimati market[3]. As for Sarlahi, even though there is little inflow to Kalimati market in the flow networks, other markets in the Kathmandu valley (belonging to Bhaktapur and Lalitpur districts) receive significant flows from Sarlahi (Figure 2f), which could, as in the previous case be routed through Kalimati market. These issues highlight some of the limitations of the gravity model, which do not account for real-world trader dynamics.

A. *Structural properties*

For each set of network parameters (β, κ, γ), there are two networks, one for each season. Both networks have 69 nodes. The cumulative distribution of flows with respect to travel time are plotted in Figures 2a–2c for different values of network parameters for

[3] http://mrsmp.gov.np/files/download/tomato%20book.pdf



(a) Sensitivity to β

(b) Sensitivity to κ

(c) Sensitivity to γ

(d)

(e)

(f)

Figure 2: Sensitivity analysis and flow validation

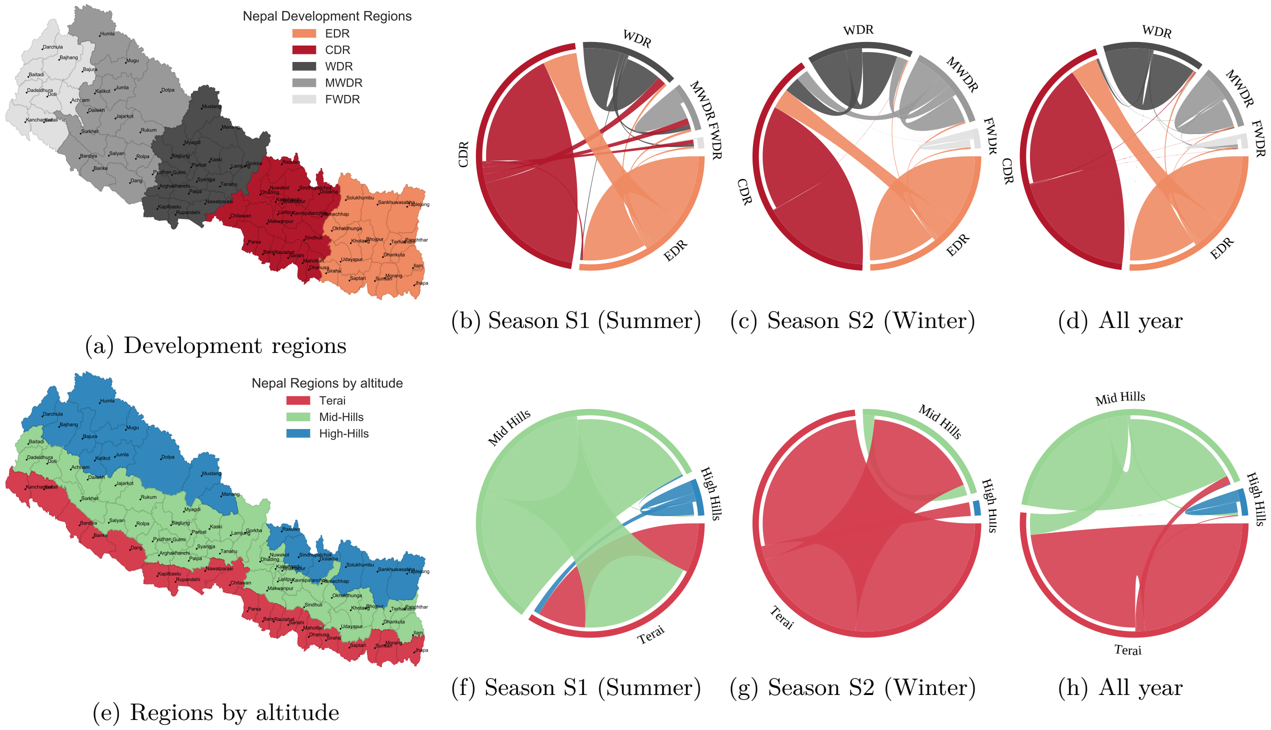
season S1 (the network corresponding to season S2 have similar properties). Except for $\beta = 1$, the flow plateaus for $t > 500$ minutes, which corresponds to ≈ 8 hours of travel time.

For further analysis of the flow network we describe the different regions within Nepal. Nepal has significant altitude variations along the North-South axis, and is divided into three major physiographic regions namely: Terai, Mid-hills and High hills (Figure 3e). For administrative reasons, Nepal has been divided along the East-West axis (Figure 3a) into five major development regions. Kathmandu, for instance, belongs to Mid-hills and Central Development Region. It is useful to remember that the Central Development Region is by far the most economically prosperous, while the population density is high along the Terai region and Kathmandu valley (Table I).

The general trends of tomato trade between markets is depicted in Figure 3 (generated for $\beta = 2$, $\kappa = 500$ and $\gamma = 1.0$). We recall that our model accounts for the fact that the Hills/Mid Hills and the Terai are the primary sources of tomato during seasons S1 and S2 respectively. This is clearly reflected in the net flow diagram between geographic regions: north (Hills/Mid Hills) to south (Terai) in S1 and south to north in S2. However, an interesting pattern to be noted is the significant flow from east to west during S1 as observed in the net flow diagram between the Development Regions. These could be due to the variability in vegetable production, and the presence of an arterial East-West highway that almost covers the entire breadth of the country.

Comparison with the annual flow network: To evaluate the importance of seasons, we constructed the annual flow network by using the gravity model with annual production for each district. The resulting flows are shown in Figures 3d and 3h. Compared to the seasonal flows we see that annual flows are of shorter distance and thus there is not much flow between regions (either between east and west or south and north).

Sensitivity analysis of the flow network: Figures 2a–2c show the sensitivity of edge weight distribution of season S1 network to β, κ and γ. We find that for $\beta \geq 2$ and $\kappa \geq 500$ the weight distribution is relatively stable. A similar behavior was observed for the season S2 flow network with respect to β and κ. Increasing γ tends to redistribute flows towards high income regions (in this case, regions around Kathmandu in the Mid Hills, Central Development Region, see Figure 3a), and leads to higher maximum flows in the network in season S1, and lower maximum flows in season S2 (not shown here). However, changing γ had minimal effect on most of the low weight edges in the network.



(a) Development regions

(b) Season S1 (Summer)

(c) Season S2 (Winter)

(d) All year

(e) Regions by altitude

(f) Season S1 (Summer)

(g) Season S2 (Winter)

(h) All year

Figure 3: **The spatio-temporal structure of the flow network.** The first row shows the flow from east to west between development regions of Nepal. The second row depicts the flow from north to south between regions of different altitudes. While the second and third columns correspond to seasonal flow, the last column corresponds to the flows generated from annual data.

B. *Role of trade network in pest spread*

We applied the network diffusion model described in Section II-C to study the role of the flow networks in the spread of *T. absoluta* in Nepal. To interpret the spread model's output, in terms of incidence reports, we need to translate t of the SI model to a real-world equivalent temporal unit (e.g., month). Validating this requires pest reports at a high spatio-temporal resolution. Since this is absent in the case of *T. absoluta*, to circumvent this problem, we make use of SI model's monotonic property: For any $t' > t$, $P_S(i, t', f_0) \leq P_S(s, t, f_0)$, and thus the ranking of relative vulnerabilities of market nodes could inform how the process unfolds. We also observed that the rank list is stable (or changes slowly) with respect to t with other parameters fixed (see Table III).

The experiment was setup under the following premise: *T. absoluta* was first introduced to the Kathmandu valley. Ground experts have high confidence in this assumption since the pest was not discovered in the previous growing season in other parts of Nepal. Given the pest reports till December 2016 (Figure 1d), we evaluate our model based on the following backward inference problem: for an observation of node states at time t, what is the most likely origin of invasion? (also known as the source detection problem [16]). We examine the likelihood of markets or regions being the source nodes, and in particular, we compare this with the likelihood of the region around Kathmandu being the source (see Figure 4). Suppose $\mathcal{O}$ is the observation criteria; it consists of pairs (v, X) where v is a node and $X \in \{S, I\}$ is a state. For each candidate initial condition f_0, we estimate the joint probability of $\mathcal{O}$ at a time step t, as a product of the marginal probability estimates from the message passing algorithm and define an *energy function* for each tuple (f_0, t) as

$$\phi(\mathcal{O}|f_0, t) = -\log\left(\prod_{(v,X)\in\mathcal{O}} P_X(i, t, f_0)\right).$$

The lower the value of ϕ, the higher the likelihood of f_0 being the initial condition. Secondly, recalling the uncertainty in interpreting time step t, we examined the relative likelihoods of each f_0 and the stability of the ranking across a range of model parameters.

We consider the spread during June-November (season S1) for model evaluation. Using the S1 flow network, our objective was to rank various starting configurations f_0 based on $\phi(\mathcal{O}|f_0, t)$ given $\mathcal{O}$, t. For a given σ, we evaluated the likelihood of each node being the central node. We considered two criteria based on which the likelihood of each f_0 as the starting configuration was computed: (i) $\mathcal{O}_{\mathrm{G}}$: this is the set of all pairs (v, I) where v is a market node that belongs to a district that reported pest presence by December 2016. (ii) $\mathcal{O}_{\mathrm{B}}$: this is the set of (v, I) for all nodes v. This is the baseline which assumes no observational data.

The results are shown in Figure 4. Firstly, we observed that for both criteria $\mathcal{O}_{\mathrm{G}}$ and $\mathcal{O}_{\mathrm{B}}$, the top few ranks are relatively robust to varying network and model parameters. Also, for both criteria, markets from the Central Development Region (CDR) that belong to Kathmandu and its adjacent districts are among the top ranked nodes. Interestingly, for the criterion $\mathcal{O}_{\mathrm{G}}$, Dhankuta (EDR), with the highest assigned production has a very low rank (Figure 4a) and a low ϕ value compared to the top market in $\mathcal{O}_{\mathrm{G}}$. However, for $\mathcal{O}_{\mathrm{B}}$, it is ranked second (Figure 4b). This clearly shows that while Dhankuta has the potential to infect a large number of areas, given what has been observed, it is very unlikely that it was the source of infection. Dhankuta reported presence of the pest only towards the end of 2016 (see Figure 1d).

Spread in season S2: To study the spread from November 2016 to May 2017, we considered the dynamics on season S2 network. To set the initial conditions, we used the results of our inference study, and chose Kathmandu with $\sigma = 10$ as the seed distribution. For this initial condition, we obtained the probability of infection for all nodes in S1 for T1 time steps. This distribution is used as initial condition for the S2 network spread. Figure 4c shows the infection probabilities for a particular combination of (T_1, T_2). As seen in Figure 4c, our model suggests that most Terai and Mid Hills regions of CDR, WDR would be affected by the end of May 2017, and subsequent seasons are only going to see increasing incidence of the pest throughout the country. From Figure 1d, we see that regions belonging to Terai in CDR and Mid Hills of WDR and MWDR have already reported pest presence (marked in Figure 4c).

While the intended usage of the origin inference formulation is to determine the source of infection, we have adapted it to compare expected spread in the model with observed data. Our results demonstrate that this framework is in general very useful in finding the likely pathways of introduction of the pest.

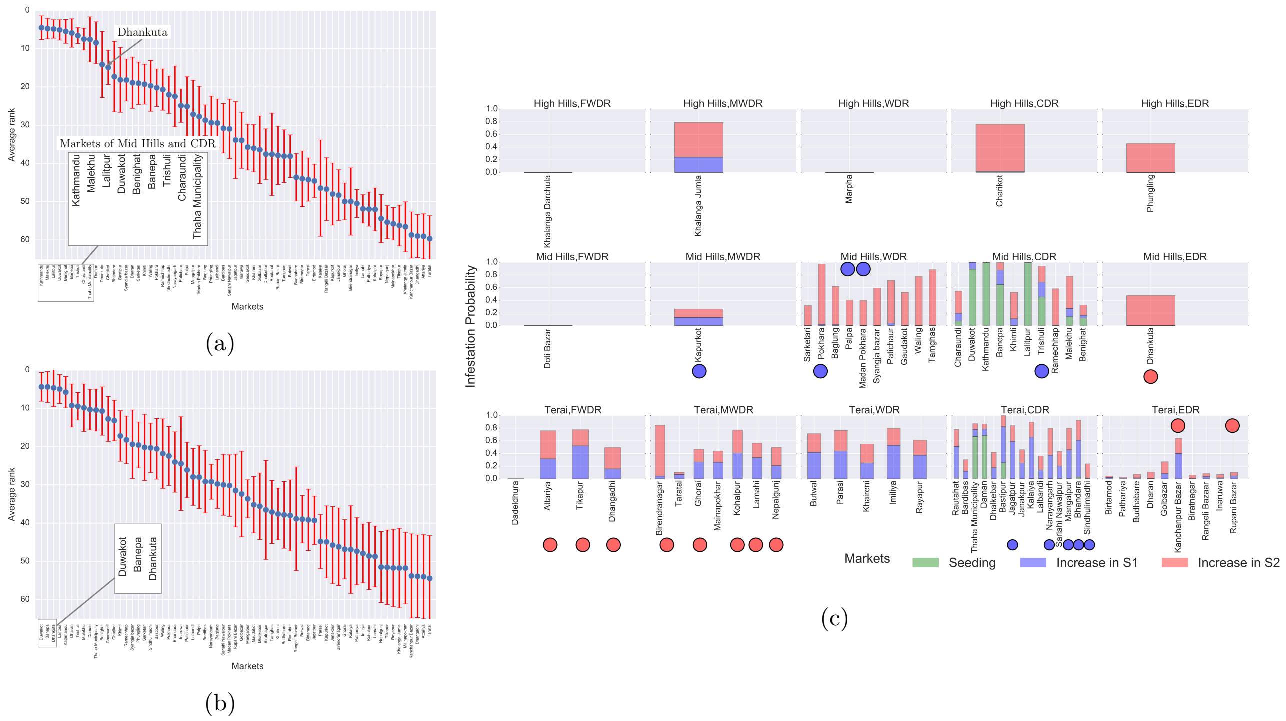
Sensitivity analyses: A full factorial design was performed with levels for the parameters of interest as given in Table III, and analysis of variance (ANOVA) was used to evaluate single parameter effect. It is worth noting that assessment of parameter sensitivity depends on the choice of quantity of interest. Since the outcome of origin inference is a ranking on markets, we used Spearman's rho to test its stability across the parameter space. The experiment was set up within the GENEUS framework [20], a general computational environment for experimental design, uncertainty quantification and sensitivity analysis.

We studied the sensitivity of individual market ranks as well as rank lists to network parameters (β, κ, γ), and diffusion model parameters (σ, t). We found that the market ranks are more sensitive to spatial seeding parameter σ and distance exponent β than other parameters. In particular, we observed that the sensitivity was highest when $\sigma = 0$ was included in the analysis. In this case (and in general for very low values of σ), substantial spread occurs only when the seed node is a source. Even if a node is in close proximity to several sources (such as Kathmandu), there is hardly any spread. This is unrealistic in the context of pest and pathogen dispersal. Hence, we restricted σ to be greater than 0 in our analysis. Also, we observe that the variance in rank is small for higher ranked nodes. This can be seen in Figure 4, and is more pronounced in the single parameter analyses. This property gives higher confidence in interpreting the results on top markets.

We used Spearman's rank correlation coefficient to analyze the rank stability. Here we use the rank list that results from configuration $(\beta = 2,\ \kappa = 500,\ \gamma = 0,\ \sigma = 5,\ T = 10)$ as the reference and calculate the Spearman's rho value with respect to it for rank lists induced by other parameter settings. Table III gives the Analysis of Variance (ANOVA) results. Under 95% confidence level, p-value < 0.05 means that the particular parameter has a significant effect. Therefore, we see that β and σ have significant effects, while others do not. Here, we note that this is despite not considering $\sigma = 0$ in the analysis.

IV. Conclusion and future work

We have described a first-principles based commodity modeling framework that integrates easily available datasets on population, production, etc.



(a)

(b)

(c)

Figure 4: **Evaluating the spread model using epidemic source inference framework.** (a) The average rank of each market based on the likelihood for the criterion $\mathcal{O}_{\mathrm{G}}$ for a range of model parameters (see Table III). (b) Same as (a), but for criterion $\mathcal{O}_{\mathrm{B}}$. (c) Spread in S2: The parameters used were $\beta = 2$, $\kappa = 500$, $\sigma = 15$, $\gamma = 1$, $T_1 = T_2 = 10$ with Kathmandu as the seed node. The blue dots correspond to markets whose districts reported *T. absoluta* presence before December 2016 (season S1), while the red dots correspond to markets which reported later.

Parm.	Levels	t-ratio	F-value	p-value
β	[0, 1, 2]	-5.16	26.6059	< 0.0001
σ	[5, 10, 15, 20]	-3.29	10.8424	< 0.0001
κ	[100, 500, 1000]	-0.42	0.1758	0.6753
γ	[0, 0.5, 1.0]	0.89	0.7970	0.3727
T	[5, 10, 20]	1.14	1.2976	0.2556

Table III: **Analyzing sensitivity to model parameters using ANOVA.**

to model the flow of agricultural produce. We have demonstrated the validity of the constructed networks, and have used it to understand the impact of commodity flow on pest spread. Despite being limited by the availability of quality validation datasets, a bare bones framework such as ours can be quickly extended to other vegetables, pests and regions with minimal effort. Our approach provides a modular framework for integration of other models that can be refined with increased availability of data and sophisticated methods.

Since our study is one of the first to consider regional commodity flow analysis in the context of pest spread, especially *T. absoluta*, there are several avenues for improvement. While some of the limitations arise from lack of refined data, others are due to the limited understanding of the underlying complexity of pest invasions. The former may be the norm for emerging contagions in a data-poor region, whereas the latter will need several iterations of model development and validation by the scientific community. Our model predominantly focuses on commodity flow, and does not explicitly account for natural or other modes of spread (infected seedlings from nurseries for example). A more comprehensive model will need to integrate ecological suitability and biology directly in the diffusion process.

Acknowledgments

This work was supported in part by the United States Agency for International Development under the Cooperative Agreement NO. AID-OAA-L-15-00001 Feed the Future Innovation Lab for Integrated Pest Management, DTRA CNIMS Contract HDTRA1-11-D-0016-0001, NSF BIG DATA Grant IIS-1633028, NSF DIBBS Grant ACI-1443054, NIH

Grant 1R01GM109718 and NSF NRT-DESE Grant DGE-154362. G.N. was also partly supported by Virginia Agricultural Experiment Station project VA-136324.

References

[1] A. S. R. Bajracharya, R. P. Mainali, B. Bhat, S. Bista, P. Shashank, and N. Meshram. The first record of South American tomato leaf miner, Tuta absoluta (Meyrick 1917)(Lepidoptera: Gelechiidae) in Nepal. *J. Entomol. Zool. Stud*, 4:1359–1363, 2016.

[2] N. C. Banks, D. R. Paini, K. L. Bayliss, and M. Hodda. The role of global trade and transport network topology in the human-mediated dispersal of alien species. *Ecology letters*, 18(2):188–199, 2015.

[3] M. R. Campos, A. Biondi, A. Adiga, R. N. Guedes, and N. Desneux. From the western palaearctic region to beyond: Tuta absoluta 10 years after invading europe. *Journal of Pest Science*, pages 1–10, 2017.

[4] L. Carrasco, J. Mumford, A. MacLeod, T. Harwood, G. Grabenweger, A. Leach, J. Knight, and R. Baker. Unveiling human-assisted dispersal mechanisms in invasive alien insects: integration of spatial stochastic simulation and phenology models. *Ecological Modelling*, 221(17):2068–2075, 2010.

[5] M. Colunga-Garcia, R. A. Haack, and A. O. Adelaja. Freight transportation and the potential for invasions of exotic insects in urban and periurban forests of the united states. *Journal of Economic Entomology*, 102(1):237–246, 2009.

[6] N. J. Cunniffe, B. Koskella, C. J. E. Metcalf, S. Parnell, T. R. Gottwald, and C. A. Gilligan. Thirteen challenges in modelling plant diseases. *Epidemics*, 10:6–10, 2015.

[7] R. Early, B. A. Bradley, J. S. Dukes, J. J. Lawler, J. D. Olden, D. M. Blumenthal, P. Gonzalez, E. D. Grosholz, I. Ibañez, L. P. Miller, et al. Global threats from invasive alien species in the twenty-first century and national response capacities. *Nature Communications*, 7, 2016.

[8] M. Ercsey-Ravasz, Z. Toroczkai, Z. Lakner, and J. Baranyi. Complexity of the international agro-food trade network and its impact on food safety. *PloS one*, 7(5):e37810, 2012.

[9] R. Y. Guimapi, S. A. Mohamed, G. O. Okeyo, F. T. Ndjomatchoua, S. Ekesi, and H. E. Tonnang. Modeling the risk of invasion and spread of *Tuta absoluta* in Africa. *Ecological Complexity*, 28:77–93, 2016.

[10] P. E. Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1):10–18, 2009.

[11] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48):1093–1103, 2010.

[12] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E*, 90(1):012801, 2014.

[13] J. F. H. Nopsa, G. J. Daglish, D. W. Hagstrum, J. F. Leslie, T. W. Phillips, C. Scoglio, S. Thomas-Sharma, G. H. Walter, and K. A. Garrett. Ecological networks in stored grain: Key postharvest nodes for emerging pests, pathogens, and mycotoxins. *BioScience*, page biv122, 2015.

[14] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[15] C. Robinet, H. Kehlenbeck, D. J. Kriticos, R. H. Baker, A. Battisti, S. Brunel, M. Dupin, D. Eyre, M. Faccoli, Z. Ilieva, et al. A suite of models to support the quantitative assessment of spread in pest risk analysis. *PLoS One*, 7(10):e43366, 2012.

[16] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Transactions on information theory*, 57(8):5163–5181, 2011.

[17] A. J. Tatem. The worldwide airline network and the dispersal of exotic species: 2007–2010. *Ecography*, 32(1):94–102, 2009.

[18] H. E. Tonnang, S. F. Mohamed, F. Khamis, and S. Ekesi. Identification and risk assessment for worldwide invasion and spread of *Tuta absoluta* with a focus on Sub-Saharan Africa: implications for phytosanitary measures and management. *PloS one*, 10(8):e0135283, 2015.

[19] USAID/Nepal. Value Chain/Market Analysis of the vegetable Sub-Sector in Nepal. 2011.

[20] S. Wu, H. Mortveit, and S. Gupta. A Framework for Validation of Network-based Simulation Models: an Application to Modeling Interventions of Pandemics. In *Proceedings of ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*. ACM, 2017.