Inferring Infection-Spreading Links in an Air Traffic Network

Lauren M. Gardner, David Fajardo, and S. Travis Waller

The objective of this paper is to present a network-based optimization method for identifying links in an air traffic network responsible for carrying infected passengers into previously unexposed regions. The required data include individual infection reports (i.e., when the disease was first reported in a region), travel pattern data, and other geographic properties. The network structure is defined by nodes and links, which represent regions (cities, states, countries) and travel routes, respectively. The proposed methodology is novel in its attempt to replicate an outbreak pattern atop a transportation network by exploiting regional infection data. The problem parallels a related problem in phylodynamics, which uses genetic sequencing data to reconstruct the most likely spatiotemporal path of infection.

The modern air traffic system provides an extensive network for human mobility and commodity exchange around the globe. However, an extensive transportation network also poses new threats to the modern world. It supersedes natural geographic barriers previously responsible for containing infectious agents and bridges previously isolated regions. As a result, both travelers and commodities are accompanied by a variety of infectious agents, such as humans and animals (e.g., insects, bacteria, and parasites), that disperse new and old diseases around the globe. The potential for virus dispersal into previously unoccupied regions has increased with the substantial rise in passenger air traffic. The burgeoning risk serves as the main motivation for this research.

The proposed problem objective is to identify the air travel routes that are most likely to be responsible for transporting infected individuals into previously unexposed regions. The network structure is representative of an air traffic system and is defined by nodes that represent regions (e.g., cities, states, countries) and links that represent air travel routes. Infection reports and properties of the air traffic network are exploited to infer the most likely spatiotemporal path of infection, which is represented as a directed spanning tree. Time-dependent regional infection counts (i.e., when and where infected individuals were diagnosed) are required to run the analysis. The limited availability of such data sets is the most significant restriction in the proposed model.

The main research contribution of this work is the use of network-based optimization methods and temporal infection data to reconstruct an infection tree atop a transportation system. Similar methodology can be applied to applications beyond disease spreading and transportation networks.

School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales 2052, Australia. Corresponding author: L. M. Gardner, l.gardner@unsw.edu.au.

Transportation Research Record: Journal of the Transportation Research Board, No. 2300, Transportation Research Board of the National Academies, Washington, D.C., 2012, pp. 13–21.
DOI: 10.3141/2300-02

LITERATURE REVIEW

The interdisciplinary nature of this work demands an understanding of transportation engineering systems, epidemiology, human mobility, human behavior, mathematics, and statistics. A comprehensive review of all such areas is beyond the scope of this paper; therefore, the literature introduced in this section highlights published work directly applicable to the problem at hand. The focus of this section is specific to related models that predict disease dispersal patterns and properties.

The introduction of mathematical models into epidemiology dates back to the early 20th century and has since been expanded on in various directions. Most ongoing epidemiological research focuses on development and implementation of agent-based simulations, which predict average disease-spreading behavioral characteristics among a population of individuals (*1–5*). These models use stochastic simulation to represent dispersal of individuals around the world and replicate the corresponding infection dynamics. In addition to the simulation-based models, analytical models that can provide quantitative measurements on the predictability of epidemic patterns have been developed to predict global epidemic behavior (*6–11*). Each of these models extends the regional-level analytical and simulation-based compartmental model to incorporate an additional scale of human mobility—specifically, interregional travel patterns. The model results are, however, speculative and predict expected outbreak behavior; they do not account for infection data or infer specific outbreak patterns. Despite their respective shortcomings, these models have proven to be invaluable in epidemiology and provide a means of predicting global epidemic patterns, comparing various interdiction scenarios (such as individual travel restrictions, city-based travel restrictions, and route-based restrictions), and conducting sensitivity analyses.

There is a significant gap in the literature for disaggregate real-time predictive and preventive measures. In particular, scenario-specific disease prediction models are of interest, since they offer many of the same benefits as the probabilistic models (e.g., aiding in strategic intervention planning) while providing a much more acute level of analysis and requiring less computational effort than the large-scale agent-based simulations.

Inferring the spatiotemporal path of a particular disease is most prevalent in the field of microbiology. *Phylogenetics* is the study of evolutionary relatedness among various groups of organisms (e.g., species, populations), discovered through molecular sequencing data. Intuitively, similar methods can be applied in predicting the spatial and temporal infection-spreading patterns of a given virus,

for which the term *phylodynamics* was coined. The main focus of phylodynamics is on developing statistical models enabling the reconstruction of timed viral dispersal patterns. For example, reconstruction of the H1N1 virus dispersal is presented by Lemey et al. (*12*), and a similar problem is presented by Wallace et al. (*13*) to infer analytically the geographic history of the H5N1 virus's migration. Cottam et al. (*14*) combine epidemiological data with genetic data to translate the genetic relatedness of pathogens isolated from infected individuals into a maximum likelihood approach to infer probable transmission trees. This is accomplished by first enumerating all possible evolutionary trees and then assigning posterior probabilities based on specifics of the mutation rates of the respective viruses. A similar analysis was conducted by Haydon et al. (*15*) with reference to the 2001 U.K. foot-and-mouth outbreak.

A novel approach to reconstruction of the spatiotemporal dynamics of outbreaks from sequence data was presented by Jombart et al. (*16*). The fundamental innovation of this approach was to seek ancestors directly from the sampled strains. The authors applied this method to track the 2009 H1N1 pandemic. The "infectious" links were selected so that the number of mutations between nodes was minimized. The problem proposed in this paper differs from the model of Jombart et al. in that the objective is to recreate the spatiotemporal dynamics of an outbreak atop a transportation network where the edges are travel routes and the edge weights are a function of regional infection data and travel patterns, rather than genetic sequencing data.

PROBLEM DEFINITION

The objective of this work is to identify the links in an air traffic network that best explain the observed regional infection data. A path of infection can be inferred between regions under the assumption that an infected individual *y* can only exist in a previously unexposed region, X, if at least one infected individual (either individual *y* or another individual *z*) traveled to X from a previously infected region at some previous time. The methodology exploits network-based optimization tools, regional infection data for a given outbreak (i.e., the day the disease "arrived" at a new location and daily infection counts), and transportation network properties (defined by the set of air travel routes and passenger volume) to infer the most likely infection tree (i.e., a set of routes on which infected individuals most likely traveled) that branches to each infected region. The goal is to identify potentially high-risk travel routes, which will aid in the development of interregional intervention strategies, security measures, and surveillance efforts. The results may also provide insight into future outbreak patterns.

ASSUMPTIONS

To implement the proposed methodology, some simplifying assumptions are necessary. Each assumption is listed and expanded on below.

1. A priori knowledge of the underlying transportation network (routes, passenger volume, travel distance, etc.) is available.
2. Temporal infection data are available for the infected regions [e.g., time of initial reported infection (i.e., time stamp) and daily infection counts].
3. A region can be infected at most once.
4. Infection spreads between regions via infected passengers traveling by air.

Information required for Assumption 1 is available from airlines as well as other private and governmental organizations. Issues concerning Assumption 2 may arise when there are multiple reported infections for a single region (which is inevitable for the state-level problem), making it difficult to identify a "time stamp" for the node, which is a necessary input for the model in identifying a causal relationship between regions correctly. There are multiple options for addressing this issue. First, on the basis of the assumption that the disease progresses within a population at a constant rate, the peak infection times (these data are available for certain diseases) can be used. Comparing these epidemic peaks would be representative of when the disease was introduced to the region. Alternatively, if a constant delay in infection reporting across states is assumed, the time of the first reported infection is a valid time stamp. The second option is chosen for the application analyzed in this paper.

Assumption 3 implicitly states that all infections in a region can be traced back to a "patient zero" within the region. Under this assumption, a region cannot be infected more than once, and there is an analogy to the susceptible-infectious-recovered model used in contact networks (which restricts an individual from being infected more than once). Assumption 3 guarantees an infection tree; that is, no cycles exist. Errors associated with this assumption could be minimized by disaggregating the problem into smaller regions (i.e., from the state to the city level), although limited availability of infection data at the city level constrains the possibility of such a model at this time.

Assumption 4 is not unrealistic for isolated states such as Hawaii or Alaska; however, in smaller, more dense regions such as the northeastern United States, many individuals travel between states by using alternative modes of transportation, and Assumption 4 is likely invalid. While air travel is assumed to be the only means of interregional travel in this model, future research plans include expanding the network structure to account for multimodal human mobility patterns. The possibility of defining a multimodal human mobility network is aided by the availability of cell phone information, currently an extensively researched topic (*17–19*).

SOLUTION METHODOLOGY

The transportation system's structure makes it an obvious candidate for network modeling tools. The network analyzed in the proposed model, $G \in (N, A)$, is defined by a set of nodes, N, which represent regions (i.e., cities, states, countries), and links, A, which connect the regions and represent air traffic patterns (i.e., travel routes). The links $(i, j) \in A$ have associated weights, $w_{ij}(\cdot)$, that are a function of node- and link-specific variables, which are discussed in detail in the section on link weights. A link weight is intended to represent the relative probability of an infected passenger traveling between a pair of regions. There is no inherent value for $w_{ij}(\cdot)$; expressions were developed to characterize the behavior of the macrolevel spreading process by using a set of variables identified as playing a role in the infection-spreading process.

The infection-spreading pattern sought forms a directed maximum probability spanning tree. Edmonds' maximum branching algorithm (*20*) was implemented on a subnetwork that includes only infected regions, $I \in N$, a feasible link set, $L \in A$, and predefined link costs, P_{ij}, which are a function of the link weights, $w_{ij}(\cdot)$, and infection time stamps, t_i. The set of feasible links $(i, j) \in L$ are those for which region i was reportedly infected before region j. It is assumed that once an infection has been reported in a region, infected individuals

continue to reside there, and that the region remains a potential source of infection to all adjacent (connected via air travel) and susceptible (uninfected) regions. Therefore, the only link feasibility constraint is $t_i < t_j$, where t_i is the time stamp for node *i*. The resultant spanning tree, *R*, includes the set of feasible links that branch to every node in *I*, such that the sum of the link costs, $\sum_{\forall(i,\,j)\in R} P_{ij}$, is maximized. Multiple link weight functions and two link cost functions were considered in the analysis.

The most computationally intensive portion of Edmonds' maximum branching algorithm is the search for and removal of cycles. The assumption that a region can be infected by at most one other region prevents the possibility of a cycle in the outbreak scenario. Therefore, implementing the algorithm for the proposed problem requires the following steps:

1. Define the set of feasible links, *L:* (i, j), where $t_i < t_j$.
2. Calculate link costs, $P_{ij}(w_{ij}(\cdot))$, for links (i, j) in feasible set *L* by using the link cost definitions in the section on link costs.
3. For each infected node, $j \in I$, select the incoming link (i, j) with the highest cost, P_{ij}, from the set of feasible adjacent links, $A|j|$, and add that link to *R*.

Steps 1 through 3 result in the maximum probability directed spanning tree, *R*.

Static Versus Dynamic Model

Two models were developed to infer the most likely infection spanning tree: static and dynamic. The main difference between the static and dynamic models is their use of infection data in the link weight function. The static model uses a single infection data point for each region (i.e., the final outbreak size), while the dynamic model uses multiple infection data points for each region (i.e., daily infection reports). (Any single point value for the infection level could be used in the static model, such as the average number of infections over the entire outbreak.) The infection data used in the static model define a single link weight for each regional pair, whereas the dynamic model defines a time-dependent link weight, $w_{ij}^t(\cdot)$, to represent the relative probability of an infected traveler entering region *j* from region *i* at time *t*. The static and dynamic model outputs are differentiated by R_S and R_D, respectively.

The following example demonstrates the main difference between the static and dynamic models discussed above. Assume that 1,000 total infection cases were reported to have occurred in Texas throughout the course of an outbreak. The static model then uses this value in $P_{ij}(w_{ij}(\cdot))$ to predict the probability that an infected traveler left Texas for, say, Ohio. In turn, Ohio reported its first case 1 week into the outbreak. However, 1 week into the outbreak, the actual number of reported cases in Texas was only 15; therefore, the link cost used in the static model (with the final infection count of 1,000) will overestimate the probability that an infected passenger arriving in Ohio (at that time) came from Texas. The dynamic model uses instead the number of reported infections in each region on each day to calculate $w_{ij}^t(\cdot)$ and thereby identifies the most likely origin of an infected passenger who arrived in Ohio on Day 7 on the basis of the number of reported infections in each previously infected region at that time.

Ignoring the progressive status of the epidemic may severely limit the predictive capability of the static model. However, the static model is a beneficial tool for prediction, useful for sensitivity analysis, and a good basis for comparison with the dynamic model. While the dynamic model may provide a more realistic prediction, the detailed data required for the dynamic model are not always available or reliable, in which case the static model could be implemented.

Static Model

For the static model, a single iteration of Edmonds' maximum branching algorithm is implemented as described in Steps 1 through 3 above, with the link costs $P_{ij}(w_{ij}(\cdot))$ and feasible link set *L* identified a priori. The algorithm identifies the incoming link with highest costs $P_{ij}(w_{ij}(\cdot))$ for each infected node. The mathematical formulation for the static model is shown below:

$$\max \sum_{\forall(i,j)\in R_S} P_{ij} x_{ij} \tag{1}$$

subject to

$$P_{ij} = f\left(w_{ij}(\cdot), t_i, t_j\right) \qquad \forall(i, j) \in R_S \tag{2}$$

$$t_i < t_j \qquad \forall(i, j) \in R_S \tag{3}$$

$$0 \le w_{ij}(\cdot) \le 1 \qquad \forall(i, j) \in R_S \tag{4}$$

$$\sum_{\forall(i,j)\in R_S} x_{ij} = |I| - 1 \tag{5}$$

$$\sum_{\forall i \in I} x_{ij} = 1 \qquad \forall j \in I \tag{6}$$

$$x_{ij} = \{0, 1\} \qquad \forall(i, j) \in R_S \tag{7}$$

$$x_{ij} = \begin{cases} 1 & \text{if edge } (i, j) \text{is in } R_S \\ 0 & \text{otherwise} \end{cases}$$

The objective requires that the set of links chosen for the spanning tree maximize the total probability of the tree. Constraints 2 through 4 pertain to the properties and dynamics of the infection process, while Constraints 5 through 7 enforce the spanning tree structure. Constraint 2 defines the link costs. The details of the link costs will be discussed in detail in the sections on link weights and link costs. Constraint 3 defines the set of feasible links—specifically, a region *i* can infect region *j* only if region *i* is infected first. Constraint 4 restricts the link weight, w_{ij}, to be fractional, a requirement due to the functional form of the link cost chosen. Constraints 5 through 7 together require that the final output be a tree; Constraint 5 requires a total of $|I| - 1$ links in R_S, Constraint 6 requires that each infected region have one incoming link, and Constraint 7 requires that the decision variable, x_{ij}, be binary.

The main concern with the static model is the implicit assumption that the probability of an outgoing traveler being infected (and thereby spreading infection into a new region) is a function of a single estimate (i.e., the final count) of infections at the route origin and not the number of infections at the time the traveler departed. This assumption would be valid if the progression of the outbreak in each region were linear; however, the objective is to predict spreading behavior at the initial stages of an outbreak, at which point this assumption is likely invalid.

Dynamic Model

The dynamic model addresses the issue of using a single point estimate by progressively building the infection tree, R_D, at single time step increments through the use of daily infection counts. The time-dependent link cost function $w_{ij}^t(\cdot)$ uses the number of reported infections in a region at time t, among other input variables, to define the relative probability of an infected passenger traveling between regions on a given day. The dynamic model is therefore able to make inferences on the basis of the real-time status of the outbreak. In the dynamic model, Edmonds' maximum branching algorithm was implemented at each time step to identify the incoming route with the highest probability of carrying an infected passenger into a newly infected region by using the time-dependent link weights, $w_{ij}^t(\cdot)$. The first iteration corresponds to the day the second region was reportedly infected. (The region with the first reported infection is by default the source node for the spanning tree.) The final iteration corresponds to the day when the last region in I was reportedly infected. The length of time between the initial reports of infection in the second and the last region is represented as T. As in the static case, the set of feasible links, L, and time-dependent link costs, $P_{ij}^t(w_{ij}^t(\cdot))$, can be calculated a priori (the set of feasible links is equivalent to the set used in the static model). The algorithm for the dynamic model is as follows:

1. Define the set of feasible links, L (this is the same set as in the static model).
2. Calculate time-dependent link costs, $P_{ij}^t(w_{ij}^t(\cdot))$, for links (i, j) in feasible set L.
3. Set $t = y$ (where y is the time stamp of the second reportedly infected region).
4. For each infected node $j \in I$ with time stamps $t_j = t$, identify the incoming link (i, j) with the highest cost, $P_{ij}^t(w_{ij}^t(\cdot))$, from the set of feasible adjacent links, $A|j|$, and add the link to R_D.
5. Set $t = t + 1$ and repeat Step 4 for T total iterations.

Steps 1 through 5 result in the dynamic maximum probability directed spanning tree, R_D. The dynamic model requires more link cost calculations a priori, but the set of feasible links is limited: for each node j the only link costs P_{ij}^t that need to be calculated are for $t = t_j$, the time step when node j was infected, and the set of links (i, j) such that $i \in A|j|$, i is adjacent to node j, and $t_i < t_j$ (region i was infected before j). The mathematical problem formulation for the dynamic model is shown below:

$$\max \sum_{\forall (i,j) \in R_D} P_{ij}^t x_{ij}^t \tag{8}$$

subject to

$$P_{ij}^t = f\left(w_{ij}^t(\cdot), t_i, t_j\right) \qquad \forall (i, j) \in R_D, \forall t \in T \tag{9}$$

$$t_i < t_j \qquad \forall (i, j) \in R_D \tag{10}$$

$$0 \le w_{ij}^t \le 1 \qquad \forall (i, j) \in R_D, \forall t \in T \tag{11}$$

$$\sum_{\forall (i,j) \in R_D} \sum_{\forall t \in T} x_{ij}^t = |I| - 1 \tag{12}$$

$$\sum_{\forall i \in I} \sum_{\forall t \in T} x_{ij}^t = 1 \qquad \forall j \in I, \forall t \in T \tag{13}$$

$$x_{ij}^t = \{0, 1\} \qquad \forall (i, j) \in R_D, \forall t \in T \tag{14}$$

$$\sum_{\forall t \in T} x_{ij}^t \le 1 \qquad \forall (i, j) \in R_D, \forall t \in T \tag{15}$$

$$x_{ij}^t = \begin{cases} 1 & \text{if edge } (i, j) \text{ is selected to be in } R_D \text{ at time } t \\ 0 & \text{otherwise} \end{cases}$$

Constraints 9 through 11 pertain to the properties and dynamics of the infection process, while Constraints 12 through 14 enforce the dynamic spanning tree structure. Constraint 15 addresses the new time-dependent decision variable, specifying that a node i can infect node j at most once over the course of the outbreak. This constraint is never violated because the algorithm used in the dynamic solution methodology evaluates each infected node $j \in I$ only once, at the time the node was first infected, $t = t_j$; because a node can be infected at most once, $\sum_{\forall t \in T} x_{ij}^t > 1$ cannot be true for any link (i, j).

Model Input

The application chosen for analysis is the 2009 swine influenza outbreak because of the availability of time-dependent infection data. The data gathered include daily volume of travel between regions (*21*), date of first reported infection in each region (*22*), daily regional infection counts (*22*), regional population (*23*), and travel distance calculated by using ArcGIS.

Network Structure

The network analyzed in this work was limited to the United States and Mexico, where the majority of infections were concentrated during the initial stages of the 2009 swine influenza. The network has 53 nodes (including the 50 U.S. states, the U.S. Virgin Islands, Puerto Rico, and Mexico) and a set of directed links connecting all the nodes having direct passenger air travel between them. Only links directed from Mexico into the United States and links between U.S. states were included in the network. No traffic exiting the United States was accounted for.

Travel volumes were aggregated to the state level to correspond to the state-level infection data set. The network was created by using the air traffic data provided by RITA (*21*), specifically passenger travel volumes aggregated across all domestic and international carriers operating within the United States. State-to-state travel volumes were calculated by aggregating passenger (airport-to-airport) travel volumes across all airports in a given state. The same aggregation was used to consolidate all travel out of Mexico into each state.

The resulting input travel data include the approximated daily passenger travel volumes from Mexico into each U.S. state and all domestic state-to-state travel in May 2009. The daily travel volume was approximated by factoring the total monthly travel volumes by 31 (the number of travel days in May). May was chosen because it was closest to the peak of the outbreak. The final network has 53 nodes (the 50 U.S. states, the U.S. Virgin Islands, Puerto Rico, and Mexico) and 1,829 links and carries more than 53 million passengers.

If city-level infection data were available for all cities in the country, the same methodology could be applied to the disaggregated city-level problem. The city-level model would likely be a better platform for

tracking the infection across space and time. The limited availability of infection data brings to light an additional motivation for this work: the need for improved infection data collection efforts and the potential benefits of making it available to researchers.

Another future research topic is introduced by the data set: Delaware has incoming flights from only one state, North Carolina, and North Carolina was reported as becoming infected after Delaware. Therefore, the model is unable to predict a predecessor for Delaware. While this issue could be attributed to incorrect data, the cause of infection in Delaware was more likely a traveler using an alternative mode of transportation. The same possibility applies to many of the nearby northeastern states, with multiple modes available for interstate travel (rail, automobile). Accounting for alternative modes of travel is a topic for future research and will be expanded on in the conclusions. In this work, the assumption that infection between states only occurs via air travel remains. Currently, Mexico is listed as the default source of infection for Delaware.

Link Weights

The model input variables listed in the section on model input are used to define the link weight function. For both the static and dynamic models, a link weight, $w_{ij}(\cdot)$, must be defined a priori because it is a necessary input for the spanning tree algorithm, and it directly determines R_S and R_D. Intuitively, the link weight should account for the interaction between two regions and the probability of passengers being infected. Specifically, w_{ij} is a function of the size of the outbreak in a given region (i.e., number of reported infections), human mobility patterns (i.e., air traffic volumes), regional populations, and travel distance. Because the historical data necessary to calibrate this type of model do not exist, the goal was instead to explore a variety of link weights that vary on the basis of their functional form and input variables and evaluate the different model outputs. The objective was to find the link weight function that best replicated the interregional infection-spreading pattern.

The various link weight functions explored, $w_{ij}(\cdot)$ and $w^t_{ij}(\cdot)$, are introduced by case number. For a given case, the only variation between the static and dynamic link weight function is the infection count variable. The same approximated daily travel volume, v_{ij}, was used for both the static and the dynamic models, and the population of a region, p_i, was assumed to remain constant over the course of the outbreak. In each of the link weight functions, the regional population size was factored by 10,000. The factor was selected so that $w_{ij}(\cdot)$ and $w^t_{ij}(\cdot)$ always fell within the range [0, 1] and was necessary because the ratio of infections to population was extremely low. The following notation defines the variables used in the remainder of this work:

$w_{ij}(\cdot)$ = weight assigned to link (i, j) for the static model;
$w^t_{ij}(\cdot)$ = weight assigned to link (i, j) for the dynamic model;
P_{ij} = cost assigned to link (i, j) for the static model;
P^t_{ij} = cost assigned to link (i, j) for the dynamic model;
T = total time span of outbreak;
t = time period (day) during outbreak;
t_i = time stamp for node i, the date of the first confirmed case in region i;
v_{ij} = approximated number of passengers traveling on route (i, j) per day;
o_i = number of reported infections in region i, used in the static model;
o_{it} = number of reported infections in region i at time t, used in the dynamic model;
p_i = population of region i; and
d_{ij} = travel distance between regions i and j.

Case Studies

The set of link weight functions represents a sample of possible expressions ranging from the simplest form, w_{ij}(I), which includes only one variable, passenger travel volume, to more complex expressions such as w_{ij}(V), derived from the binomial probability distribution, and w_{ij}(VIII), derived from the gravity model used in transportation theory. Again, the function definitions were chosen by using a set of variables expected to play a role in the spread of infection via air passengers. The analysis presented in the section on numerical results was conducted to explore the model's sensitivity to various variables and functional forms; however, the most accurate link weight function cannot be verified without link-level infection data.

In summary, the link weight functions for each case are provided in Table 1.

The explanations of each case are limited. Cases I through IV account for the proportional effect of variables in a multiplicative, single-term expression. Case V is derived by using the binomial probability distribution and defines the probability of at least one infected passenger traveling between regions. The next three cases incorporate travel distance into the link weights. Case VIII is inspired by the general gravity model used in transportation theory, which assumes that the commuting flow, f_{ij}, between regions i and j is proportional to the population in each region and the distance between the two regions.

Link Costs

The link weights, $w_{ij}(\cdot)$ and $w^t_{ij}(\cdot)$, are used in the link cost definitions, P_{ij} and P^t_{ij}, for the static and dynamic models, respectively. For the

TABLE 1 Link Weight Functions

Case	Static	Dynamic
I	$w_{ij}(\mathrm{I}) = \frac{v_{ij}}{\max_{ij} v_{ij}}$	$w^t_{ij}(\mathrm{I}) = \frac{v_{ij}}{\max_{ij} v_{ij}}$
II	$w_{ij}(\mathrm{II}) = \frac{v_{ij}}{P_i}$	$w^t_{ij}(\mathrm{II}) = \frac{v_{ij}}{P_i}$
III	$w_{ij}(\mathrm{III}) = \frac{v_{ij}}{\max_{ij} v_{ij}} * O_i$	$w^t_{ij}(\mathrm{III}) = \frac{v_{ij}}{\max_{ij} v_{ij}} * O_{it}$
IV	$w_{ij}(\mathrm{IV}) = \frac{O_i}{P_i} * v_{ij}$	$w^t_{ij}(\mathrm{IV}) = \frac{O_{it}}{P_i} * v_{ij}$
V	$w_{ij}(\mathrm{V}) = 1 - \left(1 - \frac{O_i}{P_i}\right)v_{ij}$	$w^t_{ij}(\mathrm{V}) = 1 - \left(1 - \frac{O_{it}}{P_i}\right)v_{ij}$
VI	$w_{ij}(\mathrm{VI}) = \frac{v_{ij}}{D_{ij}}$	$w^t_{ij}(\mathrm{VI}) = \frac{v_{ij}}{D_{ij}}$
VII	$w_{ij}(\mathrm{VII}) = \frac{v_{ij}}{D_{ij}} * O_i$	$w^t_{ij}(\mathrm{VII}) = \frac{v_{ij}}{D_{ij}} * O_{it}$
VIII	$w_{ij}(\mathrm{VIII}) = \frac{o_i * P_i * P_j * v_{ij}}{D_{ij}}$	$w^t_{ij}(\mathrm{VIII}) = \frac{O_{it} * P_i * P_j * v_{ij}}{D_{ij}}$

static model, the link cost function is $P_{ij} = (1 - w_{ij}(\cdot))^{(\Delta t-1)} * (w_{ij}(\cdot))$, and for the dynamic model the time-dependent link cost function is $P^t_{ij} = (1 - w^t_{ij}(\cdot))^{(\Delta t-1)} * (w^t_{ij}(\cdot))$. Both these link cost functions are referred to as *P*. In the static model, the link weight $w_{ij}(\cdot)$ is the calculated probability that infection was spread from region *i* to region *j*; therefore, $(1 - w_{ij}(\cdot))$ is equal to the calculated probability that infection did not occur between regions *i* and *j*. The exponent, $(\Delta t - 1)$, defines the number of days between the first reported infection in region *i* and the first reported infection in region *j*. The first part of the expression, $(1 - w_{ij}(\cdot))^{(\Delta t- 1)}$, therefore represents the probability that infection did not spread from region *i* to region *j* during the first $(\Delta t - 1)$ opportunities made available. The second part of the expression, $w_{ij}(\cdot)$, is the probability that infection occurred once between region *i* and region *j*. The multiplicative link cost function therefore accounts for each opportunity node *i* had to infect node *j* but did not (i.e., the number of infection delays that occurred) and the probability that infection did occur once. The same logic applies to the dynamic model. However, the link weights are recalculated each time period *t* by using the time-dependent variables.

A possible outcome from using cost function *P* that accounts for infection delay is that if a high-traffic route does not result in infection on the first opportunity made available, it is unlikely to be chosen as the cause of infection at a later time. This is because when $w_{ij}(\cdot)$ is nearly 1, the probability of the infection occurring on the second or third chance is extremely low: $(1 - w_{ij}(\cdot)) \approx 0$. This effect becomes evident when the time lapse between reported infections in two regions is large (perhaps because of faulty data or late reporting). To mitigate this issue, the methodology was also implemented by using the simple link costs $P_{ij} = w_{ij}(\cdot)$ and $P^t_{ij} = w^t_{ij}(\cdot)$, which ignore the time lapse between reported infections. These link costs functions are referred to as *w*.

When *w* is used, the model will identify more causal routes where the link weights are maximized, even if the regional time stamps are further apart. A link cost function that neglects the time gap between reported infections presents its own issues but serves as a useful tool for comparison. For both link cost functions *P* and *w*, the set of feasible links remains constant.

For simplicity, the output specific to each model (static or dynamic), case number (I through V), and link cost function (*P* or *w*) combination is represented as $R^{\text{link cost}}_{\text{model}}$ (case), or R^L_M (*C*). For example, R^W_D(V) represents the output tree for the dynamic model, Case V, with the link cost $P^t_{ij} = w^t_{ij}(\cdot)$, while R^P_D(V) represents the output tree for the dynamic model, Case V, when the link cost $P^t_{ij} = (1 - w^t_{ij}(\cdot))^{(tj-ti-1)} * (w^t_{ij}(\cdot))$.

MEASURE OF PERFORMANCE

The spanning tree for each model–case number–link cost function combination was computed for both the static and the dynamic models, for all combinations of the link cost functions (*w* and *P*) and link weight functions (Cases I through VIII), resulting in 32 spanning trees. The complete set of results is not provided, but highlights are discussed in the numerical results and analysis section.

A significant challenge associated with the proposed methodology is evaluating the model performance. The "best" model should ideally predict the actual spreading pattern that occurred, which is unknown because a complete data set of infection-causal travel routes (collected from traveler patient survey data to identify the most likely source of the disease) is not available. The lack of information makes the validity of the proposed model difficult to assess.

Comparison with Other Published Models

An alternative method of evaluation is to compare R_S and R_D with link-level predictions from other published models predicting link-level infection patterns for the 2009 swine influenza outbreak. One such model was published by Lemey et al. (*12*) and infers the phylodynamic spread in time and space of the virus by using a recently developed Bayesian statistical inference framework (*24, 25*). While the links identified by Lemey et al. do not account for factors such as human mobility or a transportation network structure, the procedure leads to a set of regional links that appropriately explain the spatial–temporal process, complemented with a formal Bayes factor (BF) test of the significance of the linkage between locations. Rates yielding a BF > 3 revealing epidemiological linkages for the United States are provided by Gardner (*26*). The strongest link is observed between Mexico and Texas. Mexico is involved in only two additional links with BF > 5, Illinois and Florida. The earliest dispersal event between Mexico and California is not supported because it precedes the time frame of the analysis. Lemey et al. (*12*) provide the only known published model that parallels the proposed model's output, and that model was therefore used as one measure against which R_S and R_D were compared.

However, relying solely on this analysis is inappropriate for multiple reasons. First, the results are probabilistic, so comparable results are not proof of accuracy, especially for the links with low BF values. Second, the pairings are not directional, so they are not equivalent to the causal links identified in the proposed model. Third, the results of Lemey et al. include only a subset of infected states because of a restricted data set and insufficient evidence for pairing two regions. Finally, the states included in the results have multiple incoming links and therefore do not form a tree structure.

Robust Analysis

Another option for evaluating the proposed model is to measure the robustness of the methodology by quantitatively comparing the output trees for the various model–case–link cost combinations. Intuitively, if many output trees share a high percentage of links, the model is robust.

R_S and R_D are constructed by selecting the incoming travel route (for each state) most likely to carry an infected individual, which forms a tree. In a tree structure, each node has a single incoming link and can be defined as a list of directed links. The percentage of links shared between two trees can be calculated easily, and a high percentage of shared links suggests that the methodology is robust with respect to variations between the two models (e.g., changes in input variables, functional form). However, a robust model does not necessarily translate to an accurate model. Accuracy of the model results can only be confirmed with link-level infection data. Although the proper data to calibrate such a model are lacking, this work is intended as a basis for developing a real-time outbreak prediction model to be implemented once more detailed infection data become available. For the remainder of this section, the "overlap" between two $R^L_M(C)$ trees is defined as the percentage of links shared.

For the cases presented, all variables except the infection count remain constant between the static and dynamic models. Therefore, the static and dynamic model outputs will correspond 100% for each case that does not include the infection variable (Cases I, II, and VI). For these cases the link weights, time stamps, and set of feasible

links are consistent between the static and dynamic models. For the remaining cases, the static and dynamic model outputs vary.

NUMERICAL RESULTS AND ANALYSIS

A comparative analysis of the case results provides insight into the role of each variable and the link weight function in the predictive capability of the model. For example, by including the regional population variable in Case II, 50% of the links selected in Case I (which only accounted for travel volume) were reassigned. As expected, the larger "from" states in Case I were replaced with less populated, highly trafficked "from" states in Case II. Similarly, a comparison of Cases I and III or Cases II and IV indicates the model sensitivity to the infection count variable. A limited overlap between the inferred spanning trees suggests that the model is sensitive to the infection count variable. Model sensitivity to the link weight function is revealed when the same set of variables is used to define two link weights, as in Cases V and VI. The highly variable but limited overlap between their output trees (as little as 13%) demonstrates the model's sensitivity to functional form.

A comparison between the pseudo-binomial Case V and the gravity-inspired Case VIII provides insight into both the role of variables and the role of the link functional form. The outputs are illustrated in Figure 1. The static case results, R_S^P (V) and R_S^P (VIII), are shown in Figures 1*a* and 1*b*, respectively; the dynamic case results, R_D^P (V) and R_D^P (VIII), are shown in Figures 1*c* and 1*d*, respectively. In the figure, each state is represented as a node, and the arrows identify the set of infectious links inferred by the model. The red nodes represent intermediate regions, which are responsible for furthering the spread of infection. While the number of shared links was numerically quantified (i.e., R_S^P (V) and R_S^P (VIII) share 37% and R_D^P (V) and R_D^P (VIII) share 9%), these illustrations bring to light characteristics of the inferred infection process and provide a way of visualizing any trends in outbreak pattern behavior. For example, certain intermediate nodes remain constant across cases, such as Washington, California, Texas, and New York. This is likely a function of these states being infected earlier in the outbreak and the increased travel volume through the states (locations of airport hubs). The ability to identify such transition locations is beneficial in developing interdiction strategy (selecting airports or regions to shut down temporarily) and in designating optimal locations for screening and surveillance systems.

Another observable trend from the illustrations is the decrease in cross-country links from Case V to Case VIII, which results in shorter link distances on average. This can be attributed to the role of distance in the gravity-inspired link weight function for Case VIII. The increased role for Texas, California, and Illinois is also evident from the figures. All three of these states ranked in the top five for infection counts and outbound travel volume. In addition,

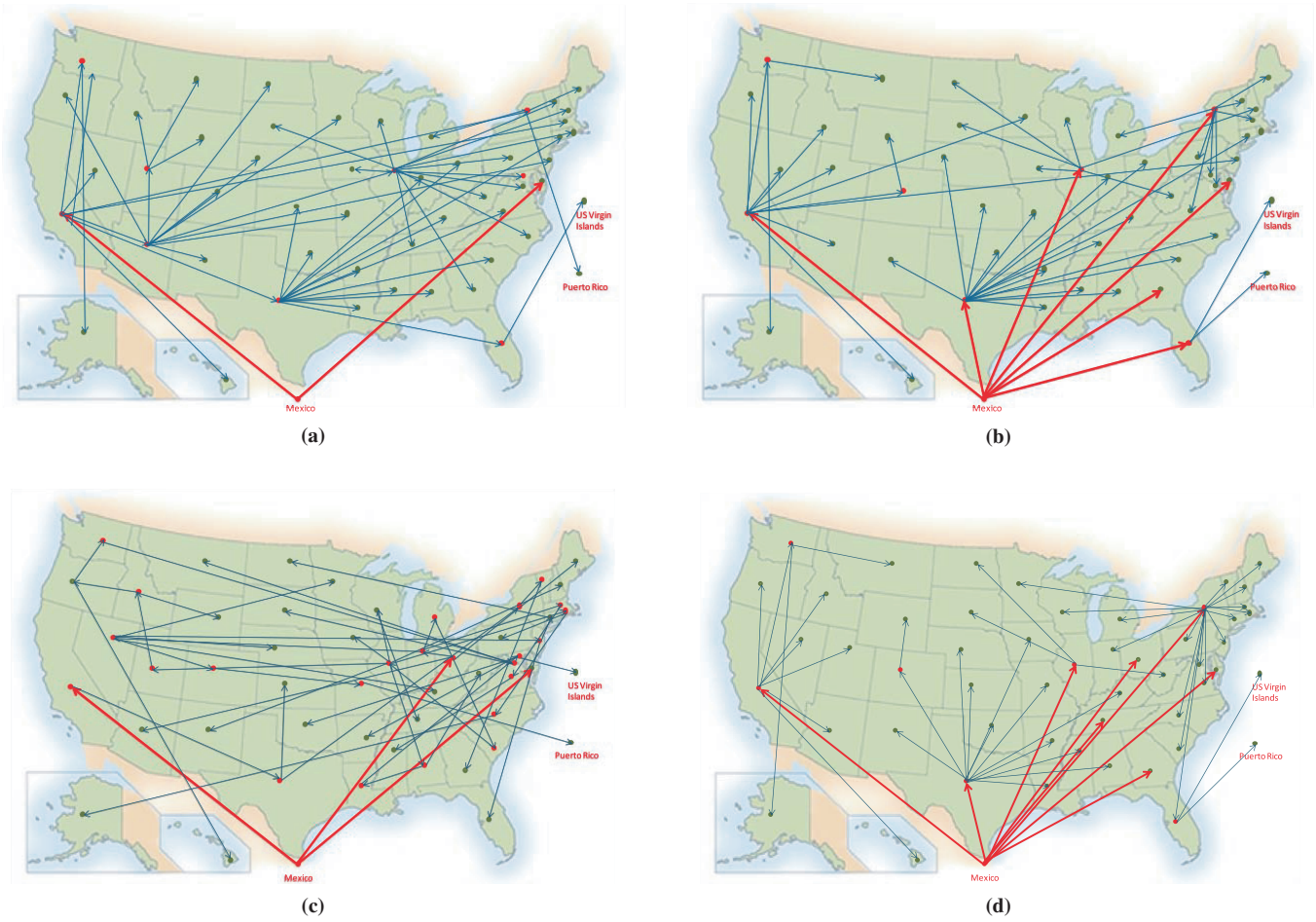

FIGURE 1 Mapped results for (*a*) Static Case V, (*b*) Static Case VIII, (*c*) Dynamic Case V, and (*d*) Dynamic Case VIII.

Texas and California were infected earliest in the outbreak, so they are feasible predecessors for most other states. The static model's predisposition to overestimate the role of regions with extremely high infection reports (at the end of the outbreak) was previously discussed and is illustrated in Case V. This effect is not as prevalent in Case VIII; the static and dynamic models identify the same set of states responsible for the majority of infections. For Case VIII the model is less sensitive to the infection variable.

The overlap between the proposed model output and phylogenetic analysis by Lemey et al. (*12*) was minimal. The lack of overlap was expected given the low confidence (BF rates) associated with the phylogenetic pairings. Seven of the 28 possible links were identified by one of the models evaluated, and in general the static model identified more of those links.

CONCLUSIONS, CONTRIBUTIONS, CRITICISM, AND FUTURE RESEARCH

The role of transportation in spreading infectious disease is of increasing importance as regional and global transportation systems expand geographically and increase in speed, efficiency, and use. The proposed modeling tool is intended (*a*) to identify the most likely air travel routes responsible for spreading disease into new, previously unexposed regions (i.e., to identify the regions adjacent to an ongoing outbreak at highest risk) and (*b*) to motivate regional-level infection data collection efforts.

Two models were introduced, static and dynamic, to reconstruct the most likely spatiotemporal path of infection atop an air traffic network. Multiple link cost functions were considered, and the associated outbreak scenario predictions were compared. The model predictions were also compared with a previously published phylodynamic analysis by Lemey et al. (*12*).

The robustness of the model to the input variables and link-based functional form was exposed by comparing the various case studies. With the current set of input data, neither the static nor the dynamic model appeared robust to variations in the link cost function and input variables. The infection data appeared to play a significant role in the model predictions, and there was little consistency between the static and dynamic models. In addition, the travel distance appeared to play a larger role in the outbreak pattern than did the regional population count. The only links that were ubiquitous across models originated in Mexico, Texas, California, or New York. This is likely a combined effect of the increased number of infections in these regions and the high travel volume out of these states. Furthermore, these regions were infected earlier in the outbreak and were therefore feasible sources of infection for a longer period of time. While these variables are all intuitive factors in the spread of infection to new regions, it is important to define a link weight function that does not overly bias the model toward these properties. With regard to the phylodynamic comparison, the proposed model results overlapped minimally with those of Lemey et al. (*12*).

While some insight can be gained by comparing the various output trees, the high level of aggregation and lack of available data made it difficult to assess the model's true prediction potential. Intuitively, the availability of dynamic infection data at the city level would improve the performance of the dynamic model over that of the static; however, it is not obvious that the dynamic model more accurately predicts the causal infection routes. This type of comparison requires further analysis and more detailed infection data.

The proposed work is more importantly a model framework, which has the potential to be expanded and applied in a much more realistic context as the necessary data become available. The novelty of the model lies in the use of infection and travel data to infer spatiotemporal outbreak patterns, which can aid in the development of real-time analysis and decision support for outbreak scenarios. The major weakness of the proposed methodology is the lack of verifiability because of limited data availability. Without link-based infection data to calibrate the model, identification of the most accurate link cost function and evaluation of certain model characteristics are not possible. In addition, the necessary high level of aggregation resulted in unrealistic assumptions and problem properties.

One major motivation of this work is the promotion of better data collection efforts. For the proposed model, collection of route-level infection data would be the most valuable. Route-level information requires data on infected individuals and their recent travel history. Such link-level infection data would permit quantitative analysis of the models' performance. In addition to link-level data, more (spatially and temporally) disaggregated infection data are necessary for implementing various proposed extensions of the models.

One obvious extension of the model is to disaggregate the problem geographically. Many of the assumptions and model properties become more realistic under a disaggregate setting, which would potentially improve the accuracy of the model. Inferring a city-to-city infection-spreading pattern is possible with the proposed methodology, but implementation is solely dependent on city-level infection data, which are unavailable.

Another planned extension of the model addresses Assumption 4—accounting for multiple modes of travel. The assumption that infections are transmitted solely via air travel is highly restrictive and likely unrealistic for many high-density areas with regions near each other. The ability to incorporate alternative modes of human transport as well as freight and cargo routes capable of transporting infectious humans (or other spreading agents) should significantly improve the model's performance. Both of these topics will be explored in depth in future research.

REFERENCES

1. Hufnagel, L., D. Brockmann, and T. Geisel. Forecast and Control of Epidemics in a Globalized World. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 101, 2004, pp. 15124–15129.
2. Grais, R., J. Ellis, A. Kress, and G. Glass. Modeling the Spread of Annual Influenza Epidemics in the U.S.: The Potential Role of Air Travel. *Health Care Management Science,* Vol. 7, 2004, pp. 127–134.
3. Cooper, B., R. Pitman, W. Edmunds, and N. Gay. Delaying the International Spread of Pandemic Influenza. *Public Library of Science Medicine,* Vol. 3, No. 6, 2006, e212.
4. Brownstein, J., C. Wolfe, and K. Mandl. Empirical Evidence for the Effect of Airline Travel on Inter-Regional Influenza Spread in the United States. *Public Library of Science Medicine,* Vol. 3, No. 10, 2006, e401.
5. Eubank, S., H. Guclu, V. S. Anil Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modeling Disease Outbreaks in Realistic Urban Social Networks. *Nature,* Vol. 429, 2004, pp. 180–184.
6. Rvachev, L., and I. Longini. A Mathematical Model for the Global Spread of Influenza. *Mathematical Biosciences,* Vol. 75, 1985, pp. 3–22.
7. Longini, I. M., Jr., P. Fine, and S. Thacker. Predicting the Global Spread of New Infectious Agents. *American Journal of Epidemiology,* Vol. 123, 1986, pp. 383–391.
8. Colizza, V., A. Barrat, M. Barthélemy, and A. Vespignani. The Modeling of Global Epidemics: Stochastic Dynamics and Predictability. *Bulletin of Mathematical Biology,* Vol. 68, 2006, pp. 1893–1921.

9. Colizza, V., A. Barrat, M. Barthélemy, and A. Vespignani. The Role of the Airline Transportation Network in the Prediction and Predictability of Global Epidemics. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 103, 2006, pp. 2015–2020.
10. Balcan, D., V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani. Multiscale Mobility Networks and the Spatial Spreading of Infectious Diseases. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 106, 2009, pp. 21484–21489.
11. Brockmann, D. Human Mobility and Spatial Disease Dynamics. In *Reviews of Nonlinear Dynamics and Complexity* (H. G. Schuster, ed.), Wiley-VCH, 2009.
12. Lemey, P., M. Suchard, and A. Rambaut. Reconstructing the Initial Global Spread of a Human Influenza Pandemic: A Bayesian Spatial-Temporal Model for the Global Spread of H1N1pdm. *Public Library of Science Currents: Influenza,* Sept. 2, 2009, RRN1031.
13. Wallace, R. G., H. HoDac, R. H. Lathrop, and W. M. Fitch. A Statistical Phylogeography of Influenza A H5N1. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 104, No. 11, 2007, pp. 4473–4478.
14. Cottam, E. M., G. Thébaud, J. Wadsworth, J. Gloster, L. Mansley, D. J. Paton, D. P. King, and D. T. Haydon. Integrating Genetic and Epidemiological Data to Determine Transmission Pathways of Foot-and-Mouth Disease Virus. *Proceedings of the Royal Society B,* Vol. 275, 2008, pp. 887–895.
15. Haydon, D. T., M. Chase-Topping, D. J. Shaw, L. Matthews, J. K. Friar, J. Wilesmith, and M. E. J. Woolhouse. The Construction and Analysis of Epidemic Trees with Reference to the 2001 UK Foot-and-Mouth Outbreak. *Proceedings of the Royal Society B,* Vol. 270, 2003, pp. 121–127.
16. Jombart, T., R. Eggo, P. Dodd, and F. Balloux. Spatiotemporal Dynamics in the Early Stages of the 2009 A/H1N1 Influenza Pandemic. *Public Library of Science Currents: Influenza,* Aug. 31, 2009, RRN1026.
17. González, M. C., C. A. Hidalgo, and A.-L. Barabási. Understanding Individual Human Mobility Patterns. *Nature,* Vol. 453, 2008, pp. 779–782.
18. Wang, P., M. C. González, C. A. Hidalgo, and A.-L. Barabási. Understanding the Spreading Patterns of Mobile Phone Viruses. *Science,* Vol. 324, 2009, pp. 1071–1076.
19. Candia, J., M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. Uncovering Individual and Collective Human Dynamics from Mobile Phone Records. *Journal of Physics A,* Vol. 41, 2008, pp. 1–11.
20. Edmonds, J. Optimum Branchings. *Journal of Research of the National Bureau of Standards,* Vol. 71B, 1967, pp. 233–240.
21. U.S. Department of Transportation, Research and Innovative Technology Administration. About RITA. http://www.rita.dot.gov/about_rita/. Accessed July 2010.
22. Centers for Disease Control and Prevention. H1N1 Flu (Swine Flu): Past Situation Updates. 2010. http://www.cdc.gov/h1n1flu/updates/. Accessed Dec. 2010.
23. United States Census Bureau. Resident Population Data: Population Change. http://2010.census.gov/2010census/data/apportionment-pop-text.php. Accessed Jan. 2011.
24. Drummond, A. J., and A. Rambaut. BEAST: Bayesian Evolutionary Analysis by Sampling Trees. *BMC Evolutionary Biology,* Vol. 7, 2007, 214. http://www.biomedcentral.com/1471-2148/7/214.
25. Lemey, P., A. Rambaut, A. J. Drummond, and M. A. Suchard. Bayesian Phylogeography Finds Its Roots. *Public Library of Science Computational Biology,* Vol. 5, No. 9, 2009, e1000520.
26. Gardner, L. *Network Based Prediction Models for Coupled Transportation–Epidemiological Systems.* PhD dissertation. University of Texas at Austin, 2011.

The Aviation Security and Emergency Management Committee peer-reviewed this paper.