

STAND: A Spatio-Temporal Algorithm for Network Diffusion Simulation

Fangcao Xu
xfangcao@psu.edu
Department of Geography,
Pennsylvania State University
State College, Pennsylvania, USA

Bruce Desmarais
bdesmarais@psu.edu
Department of Political Science,
Pennsylvania State University
State College, Pennsylvania, USA

Donna Peuquet
peuquet@psu.edu
Department of Geography,
Pennsylvania State University
State College, Pennsylvania, USA

ABSTRACT

Information, ideas, and diseases, or more generally, contagions, spread over time and space through individual transmissions via social networks, as well as through external sources. A detailed picture of any diffusion process can be achieved only when both a detailed network structure and individual diffusion pathways are obtained. Studying such diffusion networks provides valuable insights to understand important actors in carrying and spreading contagions and to help predict occurrences of new infections. Most prior research focuses on modeling diffusion process only in the temporal dimension. The advent of rich social, media and geo-tagged data now allows us to study and model this diffusion process in both temporal and spatial dimensions than previously possible. Nevertheless, how information, ideas or diseases are propagated through the network as an overall *spatiotemporal* process is difficult to trace. This propagation is continuous over time and space, where individual transmissions occur at different rates via complex, latent connections.

To tackle this challenge, a probabilistic spatiotemporal algorithm for network diffusion simulation (STAND) is developed based on the survival model in this research. Both time and geographic distance are used as explanatory variables to simulate the diffusion process over two different network structures. The aim is to provide a more detailed measure of how different contagions are transmitted through various networks where nodes denote geographic locations at a large scale.

KEYWORDS

spatiotemporal diffusion; probabilistic function; survival model; network analysis

1 INTRODUCTION

With the advent of rich social and other web-based media containing both temporal and locational information, tracing the diffusion of ideas, political opinions and even evidence of processes such as the spread of disease at a large scale has become a focus of research in recent years. Diffusion is the process by which contagions spread over space and time via complex network structures. Contagions start at specific nodes and spread from node to node over the edges of the network. These traces are called cascades.

Previous network research [7] indicates that observing when individual nodes in the network get infected by various contagions is easy, but determining the transmission pathways is difficult. In other words, the times at which nodes get infected are noted in the observational data but the sequence and parent node through

which each node gets infected is usually not. Many algorithms have been developed to simulate cascades for different contagions to drive a better understanding of diffusion process. However, most current algorithms only exploit time as an explanatory variable and don't take the geographic space into consideration.

Nevertheless, all diffusion processes that involve physical agents, locations or interactions, are embedded in geographic space. For examples, news events usually occur at specific locations. People receiving and exchanging information or ideas via social media also have a physical location, or physically being proximal to each other. When we trace the diffusion phenomena at a large scale over a long time, it has also been demonstrated the spread and adoption of many contagions are different from region to region with significant local characteristics [6, 8]. Instead of geographic distance becoming increasingly irrelevant, it's more accurate to say technology has made border less relevant.

The motivation of this research is to provide a well-defined and mathematically solid approach for solving diffusion simulation and modeling problems taking both spatial and temporal information into consideration. To achieve this, we have developed a probabilistic algorithm called STAND, using a survival approach, to simulate spatiotemporal diffusion cascades. This algorithm is applicable to various types of network structures. It is intended that our research can lead to new insights of how different contagions, including topics, opinions, sentiments, or events mined from world wide web are propagated over space and time.

2 RELATED WORK

Simulating the complete topology of spatiotemporal networks and multiple contagions spread over them is challenging for two reasons: First, in many cases we can only observe the timing information of when nodes get infected [11]. Second, even though large amounts of digital heterogeneous data are now available via the World Wide Web, locational information is extremely sparse, unstructured and often ambiguous [1, 4, 13]. Developing a flexible model for deriving the network structure and cascade behaviors is key to uncovering the mechanism that governs spatiotemporal diffusion processes and their dynamics.

Several differential equation models (DE) and agent-based models (AB) have been proposed to simulate the network structure and spatiotemporal diffusion cascades. DE models [9, 10] usually aggregate agents into several states (e.g., infected or uninfected). The transitions among different states are modeled by differential equations. In contrast, agent-based models [2, 5] have considered the heterogeneity of agents. They simulate the diffusion in realistic networks by defining how the infection may occur through

Table 1: Parameters in Diffusion Networks

Parameters	Def
$G(V, E)$	Directed Graph with node set V and edge set E
c	Contagion that spreads over G
C	Set of contagions c
T	Cascade propagation tree
$T_c(G)$	Set of all possible cascade trees of the contagion c
t_i	Time when node i get infected by a contagion
$\Delta_{i,j}$	Time difference between the node infection time $t_j - t_i$
α	Diffusion speed scaling parameter
β	Probability that contagion spreads over the edge of G

individual-based interactions (e.g., infection can only occur when agents are at the same location).

While, most existing simulation models of network diffusion are based on assumptions of agent homogeneity or how agents interact with each other for spreading contagions, none have integrated space and time together to account for diffusion probability individually for each node-pair over spatiotemporal networks at large-scale. We also intend to fill this gap. We develop an exponential diffusion probabilistic algorithm that integrates geographic distance, node infection time and transmission speed together based upon NETINF [11]. The reasons to choose an exponential network diffusion model as a starting point for STAND are: 1) it's a continuous-time model without any assumptions about the individual interactions, 2) it is amenable to inference/estimation for different datasets, and 3) it has a flexible structure in terms of adding features like geo-distance. Other time distributions (e.g., log-normal, Gamma, etc.) can also be considered in the future research as well.

3 THE STAND ALGORITHM

Table I defines the parameters used in many diffusion network models, and we build upon this notation. When a specific contagion spreads over the network, it will create a cascade by infecting nodes in a temporal sequence. The cascade of a contagion c is denoted by a directed tree T , consisting of a set of infected nodes with observed infection time: $(i, t_i)_c$ where $i \in V$.

3.1 Temporal Probabilistic Survival Likelihood

The likelihood of a contagion c spreading over an edge (i, j) in NETINF is calculated by a probabilistic exponential function and assumed to depend only on the time difference $\Delta_{i,j}$ in Equation (1).

$$P_c(i, j) \approx P_c(\Delta_{i,j}) \propto \frac{\beta}{e^{\alpha \Delta_{i,j}}} \quad (1)$$

Only the first time when a node gets infected will be counted for each contagion c and the contagion must diffuse forward in time. If there is a node j that never got infected by any contagion, then $t_j = \infty$ and $\Delta_{i,j} \approx \infty$. Thus each infected node will only have one parent node who diffuses the contagion c and the $t_j > t_i$.

However, there may exist more than one possible diffusion cascade tree for the same set of infected nodes. These trees have the same temporal sequence for the occurrences of infections among nodes but their diffusion paths are different. Equation (1) which only exploits the infection time difference to explain the diffusion probability, cannot solve which diffusion path has a higher probability.

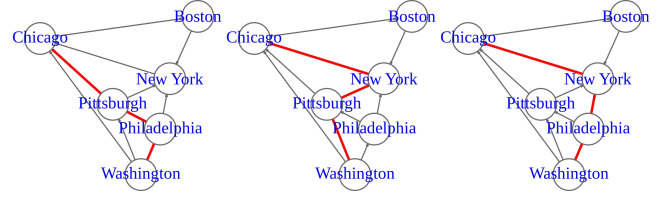


Figure 1: Different cascade trees

For example, Figure 1 shows three different trees with the same observed spatiotemporal diffusion sequence where the nodes are geographic locations, $(t_{Washington}, t_{Philadelphia}, t_{Pittsburgh}, t_{Chicago})_c$. The infection time $t_{Washington} < t_{Philadelphia} < t_{Pittsburgh} < t_{Chicago}$. The diffusion path from Washington to Chicago can be $(Washington \rightarrow Philadelphia \rightarrow Pittsburgh \rightarrow Chicago)$ as shown by the red line in the left plot or $(Washington \rightarrow Pittsburgh \rightarrow Chicago)$ in the middle plot.

The diffusion probability from Washington to Chicago of these three paths defined by Equation 1 are same, demonstrated as below:

$$\begin{aligned} P_c(i, k, p, j) &= \frac{\beta}{e^{\alpha \Delta_{i,k}}} * \frac{\beta}{e^{\alpha \Delta_{k,p}}} * \frac{\beta}{e^{\alpha \Delta_{p,j}}} = \frac{\beta^3}{e^{\alpha(\Delta_{i,k} + \Delta_{k,p} + \Delta_{p,j})}} \\ &= \frac{\beta^3}{e^{\alpha \Delta_{i,j}}} = \frac{\beta^3}{e^{\alpha(\Delta_{i,p} + \Delta_{p,q} + \Delta_{q,j})}} = P_c(i, p, q, j) \\ &= \frac{\beta^3}{e^{\alpha(\Delta_{i,k} + \Delta_{k,q} + \Delta_{q,j})}} = P_c(i, k, q, j) \end{aligned} \quad (2)$$

where i is Washington, k is Philadelphia, p is Pittsburgh, q is New York and j is Chicago.

It's thus needed to extend the current diffusion probabilistic algorithms to include space as well as time to better model the dynamic diffusion over the spatiotemporal network.

3.2 Spatiotemporal Probabilistic Survival Likelihood

We assume the spatiotemporal diffusion probability $P_c(i, j)$ that a cascade c will spread from a node i to a node j decreases with both the spatial distance d_{ij} and time difference $\Delta_{i,j}$ in a certain way. Geographic distance is integrated into Equation (1) as a starting point. Considering that the geographic distance may have direct influence on the diffusion speed and infection time interval, this influence is modeled by the survival analysis model.

Supposing T is the observed infection time of a node, then the probability that this node would not yet be infected at any time t is denoted by the survival function, $S(t) = P(T > t)$.

The probability that a given node would get infected at any time t , is denoted by a cumulative probability density function (CDF), $F(t) = P(T \leq t) = 1 - S(t)$. The probability that a node will get infected within a time interval $(t, t+dt)$ is $P(t \leq T \leq t+dt) = f(t)dt$ where $f(t)$ is the infection rate over time, given by the probability density function (PDF):

$$f(t) = \frac{d}{dt} F(t) = \frac{d}{dt} (1 - S(t)) = -\frac{d}{dt} S(t) \quad (3)$$

The instantaneous infection rate $h(t)$ at the given time t , is called the hazard rate, that reflects the likelihood that the uninfected node will get infected within a very short time interval $(t, t + dt)$, given this node hasn't been infected before:

$$h(t) = \lim_{dt \rightarrow 0} \frac{Pr(t \leq T < t + dt)}{dt \times S(t)} = \frac{f(t)}{S(t)} \quad (4)$$

The survival likelihood $S(t)$ and the cumulative infection probability $F(t)$ have the following relationships with the hazard rate $h(t)$:

$$S(t) = e^{-\int_0^t h(t)dt} \quad F(t) = 1 - e^{-\int_0^t h(t)dt} \quad (5)$$

where the $\int_0^t h(t)dt$ is the cumulative hazard that represents the total risks of a node being infected up to the time point t .

In order to measure pairwise infection likelihood in Equation (1), the probability density $f(t)$ and the hazard rate $h(t)$ are needed to calculate the probability that a node j gets infected by the node i within the time interval $(t_i, t_i + dt)$. A higher value of density function $f(t)$ means the node j is more likely to be infected by the node i within a short time interval dt which results in a higher hazard rate.

In this step, we use the a proportional hazard function suggested by [12] to exploit the geographic properties in the form below:

$$h(t; Y; \lambda; \theta) = e^{(\lambda X + Y)} h_0(t; \theta) \quad (6)$$

where the $h_0(t; \theta)$ is the baseline hazard function, and t is the observed infection time. X is a vector of explanatory variables associated with node properties and Y is a vector of explanatory variables associated with the spatial dependencies. λ and θ are parameter coefficients for the explanatory variables X and the baseline hazard function.

To simplify, the baseline hazard rate $h_0(t; \theta)$ in Equation (6) is assumed to be constant and independent with time for any node-pairs. Thus the $h_0(t; \theta)$ and its associated cumulative hazard rate $H_0(t; \theta)$ are:

$$h_0(t; \theta) = \theta \quad H_0(t; \theta) = \int_0^t h_0(t; \theta)dt = \theta t \quad (7)$$

Furthermore, only the spatial distance is implemented into the diffusion probability as a starting point. Thus the λX is a constant modeled by the parameter λ and the explanatory variables Y , accounting for spatial dependencies in Equation (6) is modeled by the distance between any node-pair i and j with a decay parameter. To unify the parameter coefficients, the λ for node properties is reset as λ_0 and the decay parameter for spatial dependencies (i.e., distance here) is set as λ_1 . Equation (6) can be reformatted as:

$$h(t; Y; \lambda; \theta) = \theta e^{(\lambda_0 + \lambda_1 d_{i,j})} \quad (8)$$

where the $d_{i,j}$ is the geographic distance between the node i and j .

The cumulative probability $F(t)$, the probability density $f(t)$, and the infection likelihood defined in Equation (1) with the hazard rate $\theta e^{(\lambda_0 + \lambda_1 d_{i,j})}$ over the infection time interval $\Delta_{i,j}$ can be rewritten

as:

$$\begin{aligned} F_{i,j}(\Delta_{i,j}) &= 1 - e^{-\int_0^t h(t)dt} = 1 - e^{-\theta e^{(\lambda_0 + \lambda_1 d_{i,j})} \Delta_{i,j}} \\ f_{i,j}(\Delta_{i,j}) &= \frac{\theta e^{(\lambda_0 + \lambda_1 d_{i,j})}}{e^{\theta e^{(\lambda_0 + \lambda_1 d_{i,j})} \Delta_{i,j}}} \\ P_c(i, j) &\propto \beta f_{i,j}(\Delta_{i,j}) = \frac{\beta \theta e^{(\lambda_0 + \lambda_1 d_{i,j})}}{e^{\theta e^{(\lambda_0 + \lambda_1 d_{i,j})} \Delta_{i,j}}} \end{aligned} \quad (9)$$

The network inference problem can also be solved by developing a greedy algorithm to search for all edges that can maximize the probability of a network structure over which all contagions C can spread.

4 EXPERIMENTAL SIMULATION

In order to simulate real-world diffusion processes, and evaluate our algorithm for recovering the network structure on the synthetic cascades, experiments are described in the following sections.

4.1 Simulation of Ground-truth Network

One hundred geographic cities/counties with the largest population in the United States, excluding Hawaii and Alaska, are taken as spatial network nodes V . The original data is from the U.S. Geological Survey collected in 2017. Some cities/counties having high population and very close geographic locations were merged into larger metro areas (e.g., New York City or Greater Los Angeles). Then we manually deleted a few small cities and add other cities in Montana, Wyoming, North and South Dakota to make the nodes more evenly distributed over the United States.

Considering that many real-world networks share some similar properties, such as relatively small average path length, high clustering coefficient, or high reciprocity (i.e., the percentage of mutually connected node-pairs), two directed network structures are considered to generate the ground-truth network G . They are the random network structure [3] and the small-world network structure [14]. Characterizing how contagions spread over these two extreme network structures can help us better understand more complex network diffusion phenomena. However, our proposed STAND algorithm can be applied to any kind of network structure. The pseudocode of STAND algorithm is given as below:

4.1.1 Simulation of Random Network. The random network assumes the node degree in the network has a normal distribution and edges are randomly picked with a constant probability with a same node-edge density. This has been widely used for network simulations with the absence of topological information about the network structure.

Building this type of network begins with n spatial isolated nodes and selecting out of those any two to randomly place an edge between them until all edges of the required number have been added to the network without repetition. An example is shown in Figure 2 (Left).

4.1.2 Simulation of the Small-world Network. The small-world network as shown in Figure 2 (Right) assumes the node degree follows the power-law distribution, in which only a small proportion of nodes has a large number of links while the majority has only few or no links. Thus it has a high clustering coefficient and

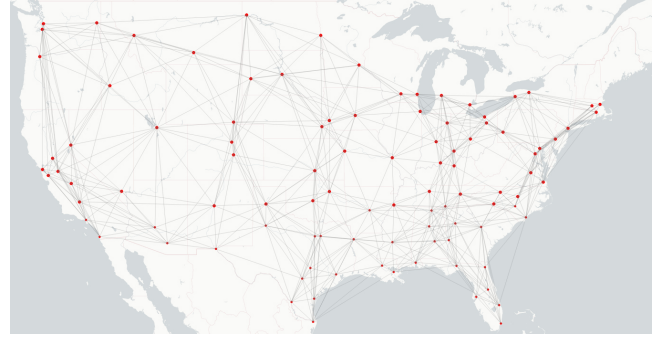
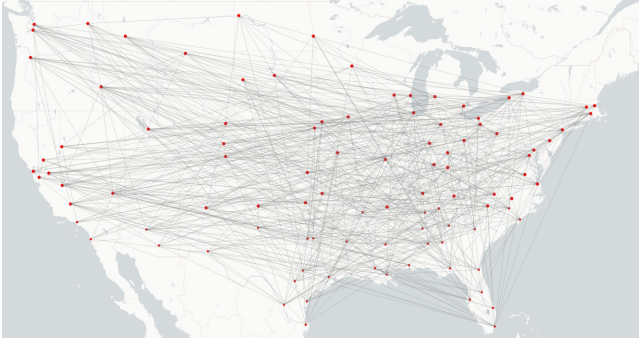


Figure 2: Random network (Left) and Small-world network (Right) with 100 nodes and 600 edges.

Algorithm 1 STAND Algorithm

Require: coordinates of n nodes (lng, lat) , number of edges m , network $G(V, E)$, parameter $\beta, \theta, \lambda_0, \lambda_1$, cutoff diffusion time t_{max} , number of nearest neighbours k .

Step 1: Simulate Random & Small-world Networks

$V \leftarrow (lng, lat)$

$G_{random} \leftarrow E \leftarrow COMBINATION(V, 2, m)$

for all node $v \in V$ **do**

$G_{sw} \leftarrow E \leftarrow (v, K_NEAREST(v, k, Euclidean))$

$RECONNECT(E \in G_{sw}(V, E), 0.05)$

return G_{random}, G_{sw}

Step 2: Simulate Cascades

for all edges $(i, j) \in G_{random} | G_{sw}$ **do**

$d_{i,j} \leftarrow EuclideanDist((lng_i, lat_i), (lng_j, lat_j))$

$h_{random} | h_{sw} \leftarrow h(i, j) = \theta e^{(\lambda_0 + \lambda_1 d_{i,j})}$ Equation (8)

return h_{random}, h_{sw}

$C_{random} = SIMULATE(G_{random}, h_{random})$ Equation (9)

$C_{sw} = SIMULATE(G_{sw}, h_{sw})$

return simulated cascades C_{random}, C_{sw}

a relative short average path length. The small-network model is widely used in many application contexts, such as the outbreak of disease or social activities.

Building this type of network starts with connecting each spatial node to its k nearest neighbors. Next, a small proportion of edges are randomly rewired by removing the original edge and reconnecting the starting node of the original edge to another node with a probability p (i.e., 0.05).

Here, we have built 600 edges among these geographic locations for each network structure. In comparing left and right figures in Figure 2, it is obvious that the average edge distance of the random network is much longer than that of the small-world network, which has made the small-world network seem sparser over space than the random network even though they have the same number of edges. The small-world network also shows a significant clustering pattern of cliques in contrast to the random network.

Table 2: Parameters of λ_1 and t_{max}

λ_1	$-5 * 10^{-5}$	$-3 * 10^{-5}$	$-2 * 10^{-5}$	-10^{-5}
	$-9 * 10^{-6}$	$-8 * 10^{-6}$	$-7 * 10^{-6}$	$-6 * 10^{-6}$
	$-5 * 10^{-6}$	$-4 * 10^{-6}$	$-3 * 10^{-6}$	$-2 * 10^{-6}$
t_{max}	500	1000	2000	4000
	8000	∞		

4.2 Simulation of Cascades

A set of cascades C that spreads over the network G are generated by the probabilistic function defined in Equation (9). At this step, each node is selected as the single starting infected node with assigned an infection time 0, and then it starts to spread the contagion to every remaining node. The infection time of remaining nodes is calculated based on Equation (9). If there is no edge between any two nodes, the hazard rate α is forced to be set as 0.

Different parameter values are selected for generating diffusion cascades. Recalling that the hazard rate $\alpha = \theta e^{(\lambda_0 + \lambda_1 d_{i,j})}$ in Equation (8) controls the infection rate that a contagion spreads over edges, there are three parameters to express this: θ (i.e., a baseline hazard rate), λ_0 (i.e., influence of node properties), and λ_1 (i.e., parameter coefficient of influence of distance).

Several parameters are set as fixed values in this step because they are constant scaling factors irrelevant to account for the influence of explanatory variables (i.e., the distance). They are: 1) the prior probability of an edge to successfully spread the contagion, $\beta = 0.5$. A higher value of β means most of edges are more likely to spread the contagion, which results in large infections/cascade size within the network; 2) the baseline hazard rate, $\theta = 0.5$; 3) the influence of node properties on the diffusion hazard rate, $\lambda_0 = 1$; and 4) the penalty parameter that accounts for the small probability that the contagion may spread via non-existing edges, $\epsilon = 10^{-4}$.

The parameter coefficient of the influence of the distance λ_1 and the cutoff diffusion time t_{max} used to decide to what extent the diffusion should fail when the diffusion time along an edge is too long, are chosen from a list of values in Table II.

These values of λ and t_{max} increase monotonically and can represent cascades ranging from small (i.e., contagions can almost not spread) to large (i.e., a large infection over all nodes). The

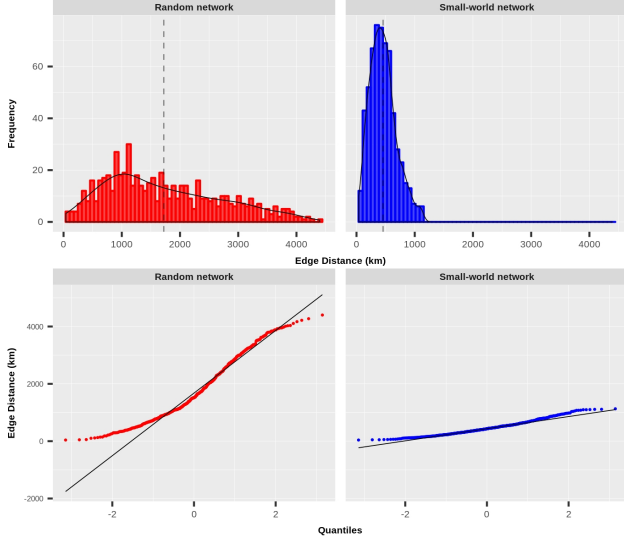


Figure 3: Histogram and QQ-plot of the distance between any two nodes in the Random and Small-world networks

reason to set λ_1 as negative is that in general, the hazard rate α should be small when the spatial distance is very large. In other words, the diffusion time that a contagion spreads via a network edge is longer when two nodes are far away from each other if all other parameters are fixed. The α has a positive relationship with λ_1 . When λ_1 increases from the $-5 * 10^{-5}$ to $-2 * 10^{-6}$, the diffusion/infection speed via network edges will be larger.

The cutoff diffusion time t_{max} is used to set those nodes that are infected beyond the threshold as ∞ . For example, if 1000 (time units) is selected as the cutoff value, then nodes with simulated infection time more than 1000 should be set as "won't get infected". It can model phenomena that the outbreaks of some contagions happen within a short period then fades very quickly due to geographic constraints while others can easily spread at a large scale and last for a long time.

The analysis of the simulated cascades of the static random network and the static small-world network is given in the following section using the parameters selected above.

5 ANALYSIS OF RESULTS

5.1 Distribution of the Distance

The histograms and Quantile-Quantile (QQ) plots in Figure 3 are used to examine the statistical distribution of the edge distance. The range of the statistical distance (km) between two nodes in the random networks is much larger than that in the small-world network. Both of them have a bell-shaped curve as shown in the top of Figure 3. However, the density curves in black suggest that the distribution of the edge distance in the random network is smoother and more widely distributed while the edge distance the small-world is dominated by more lower values. The reason accounting for this in the random network may be due to the limited number of geographic points, thus few observations can easily skew the distribution. For the small-world network, the reason is the existence of a small portion of edges that connect different cliques.

Table 3: Information about Simulated Cascades

node_name	time	parent_node	v_path	cascade_id
3: Chicago	0.00000	3	3	3
31: Milwaukee	1.71148	3	(3, 31)	3
56: Fort Wayne	3.61669	3	(3, 56)	3
82: Grand Rapids	4.03627	56	(3, 56, 82)	3

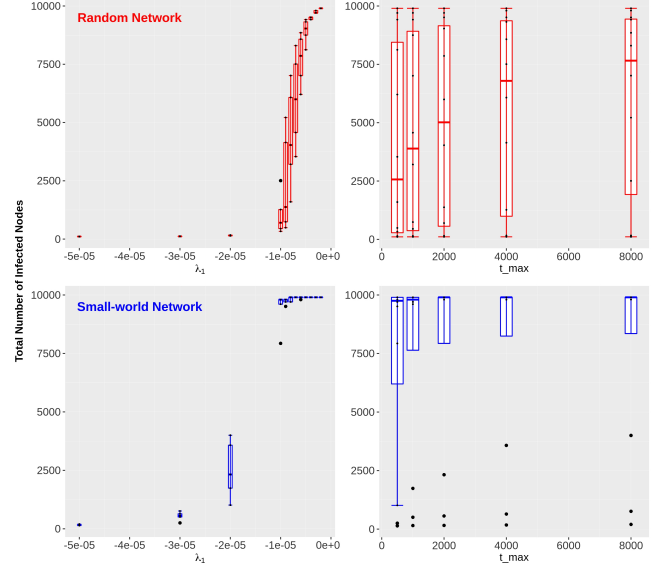


Figure 4: Box plots of the number of infected nodes in 100 simulated cascades for Random network and Small-world network

5.2 Infected Nodes in Simulated Cascades

For each combination of λ_1 and t_{max} shown in Table II, 100 cascades are generated by taking every node as a starting infected node and the information including the infected node name, the infection time, the parent node from whom the contagion spreads, the diffusion pathway and the cascade id are reported. For example, the information in Table III suggests a diffusion cascade that starts from the node Chicago spreads via two pathways (*Chicago* \rightarrow *Milwaukee*) and (*Chicago* \rightarrow *Fort Wayne* \rightarrow *Grand Rapids*).

The box-plots in Figure 4 give a clearer picture of the distribution of the number of infected nodes under different sets of λ_1 and t_{max} . The left series of plots in Figure 5 shows how the total number of infected nodes changes for different values of λ_1 , fixing the cutoff diffusion time while the right plot in Figure 5 shows how the total number of infected nodes changes for different cutoff diffusion time t_{max} , fixing λ_1 .

Several conclusions about the influence of λ_1 and t_{max} and the difference between the random network structure and the small-world network structure are drawn from these statistical results:

- (1) If the cutoff diffusion time is set as ∞ where all nodes can get infected (i.e., the bottom row of the left plot in Figure 4, the value of λ_1 has no influence on the total number of infected nodes;
- (2) For other cutoff diffusion times, it is clear that the geographic distance has a great influence on the total number

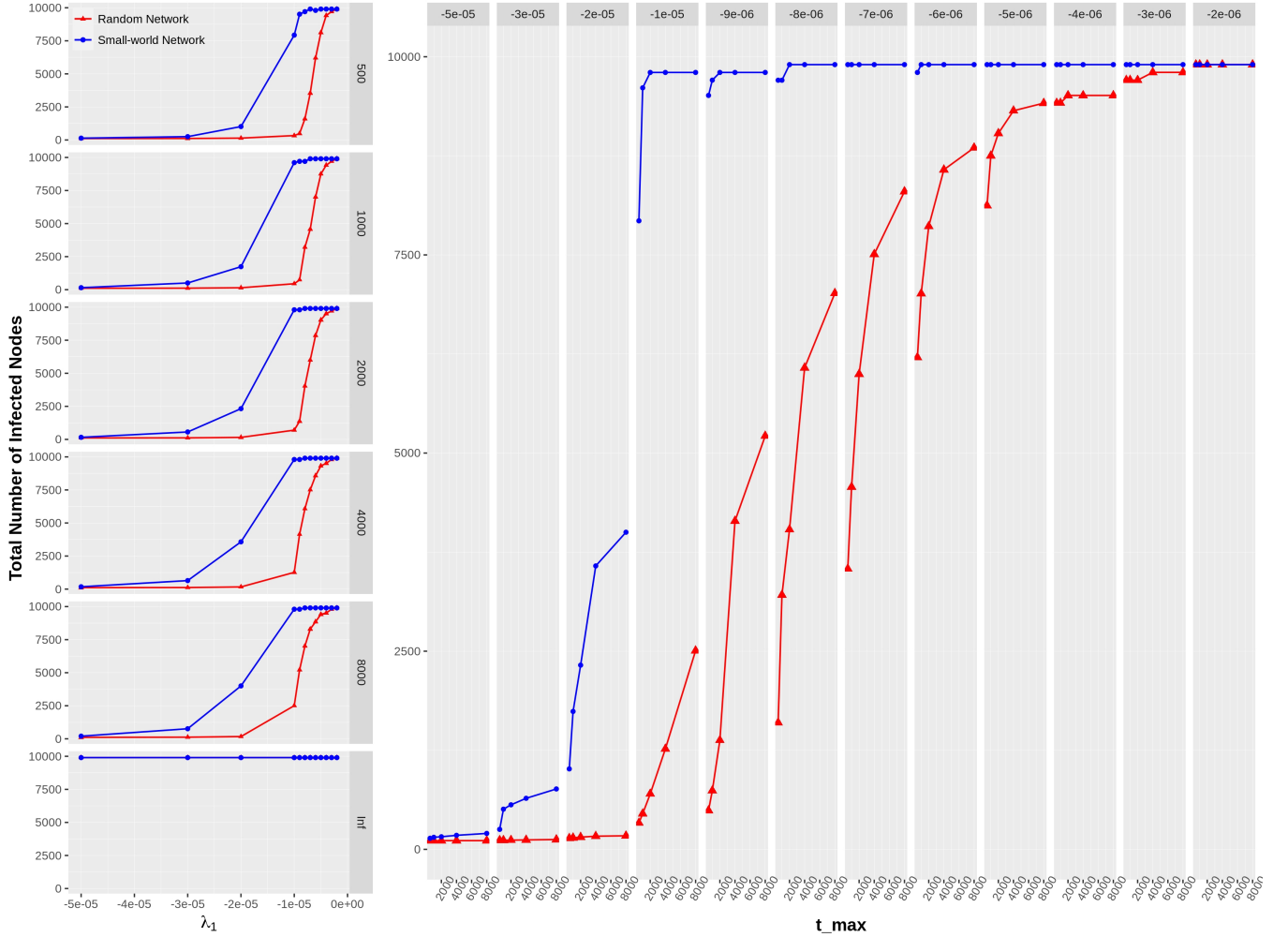


Figure 5: Number of infected nodes in 100 simulated cascades for the random networks and small-world networks

of infected nodes. The number of infected nodes monotonically increases to the peak when λ_1 increases;

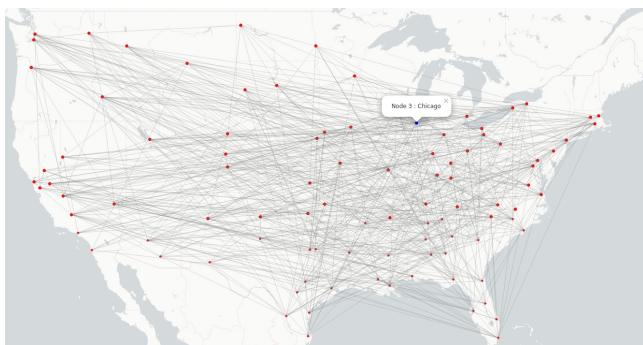
- (3) When the **absolute value** of λ_1 is low, the influence of the distance on the diffusion is small, which results in an approximate constant hazard rate for all location-pairs;
- (4) When the **absolute value** of λ_1 is high, more nodes will survive from contagions due to the geographic constraint as first several columns of the right plot in Figure 4;
- (5) The maximum number of infected nodes that the small-world network can reach is more than that in the random network because the small-world network is highly clustered that makes the contagion spreads more easily;
- (6) The simulation process of the random network is more sensitive to the change of parameters and the number of infected nodes will increase gradually when λ_1 and t_{max} increase. This number in the small-world network will increase dramatically to the peak within a short range of λ_1 and is less sensitive to the change of the cutoff diffusion time except when $\lambda_1 = -2 * 10^{-5}$.

When λ_1 is very small, most nodes fail to spread contagions for different cutoff diffusion time except for ∞ . When λ_1 increases, most nodes in the small-world network can get infected within a short time due to an average small distance. So the influence of the cutoff diffusion time is only obvious for λ_1 around $-2 * 10^{-5}$. The distribution of the number of infected nodes of the random network is nearly normally distributed while that of the small-world network is more skewed with a large proportion of outliers identified.

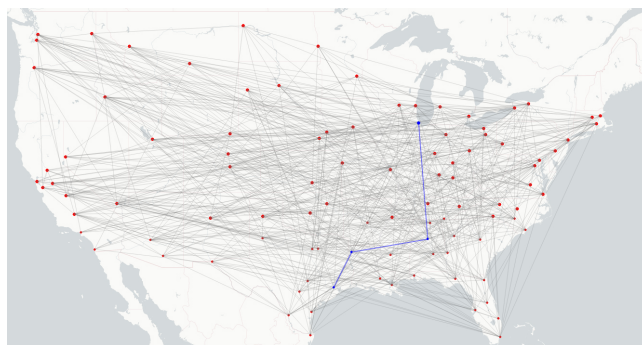
6 VISUALIZATION OF DIFFUSION CASCADES

To better understand the geographic distribution of the infected nodes in the diffusion process, Figure 6 below shows simulated cascades generated by taking **Chicago** as the starting infected node under different sets of λ_1 and cutoff diffusion time t_{max} for the random network and the small-world network.

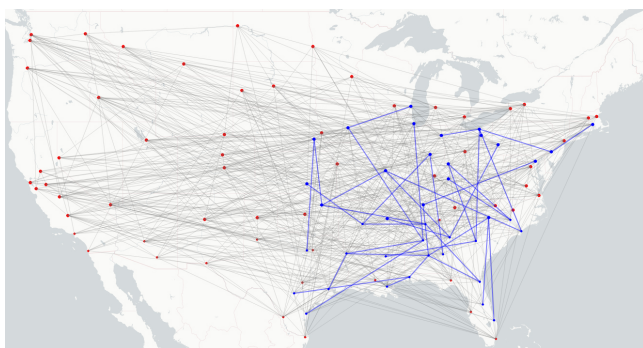
When λ_1 and cutoff diffusion time t_{max} is very small, it is hard for contagions to spread over the random network. However, contagions can still spread from Chicago to Milwaukee in the small-world network when $\lambda_1 = -5 * 10^{-5}$ and $t_{max} = 500$ due to nodes in



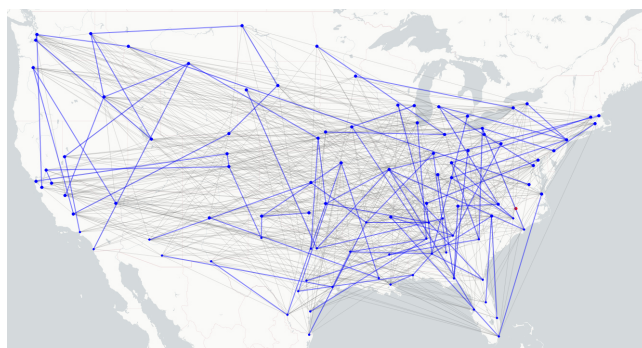
$$\lambda_1 = -3 * 10^{-5}, t_{max} = 1000$$



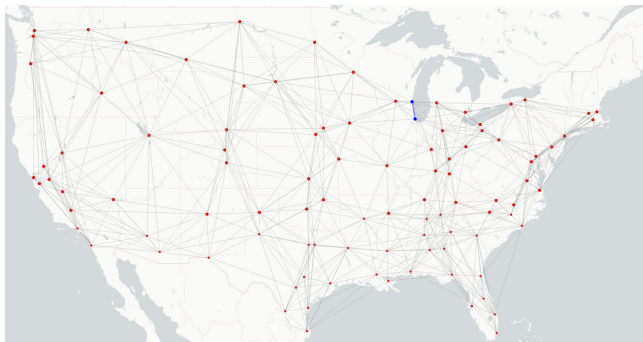
$$\lambda_1 = -10^{-5}, t_{max} = 1000$$



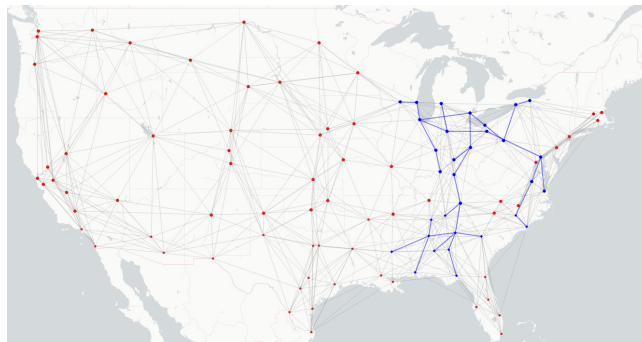
$$\lambda_1 = -10^{-5}, t_{max} = 4000$$



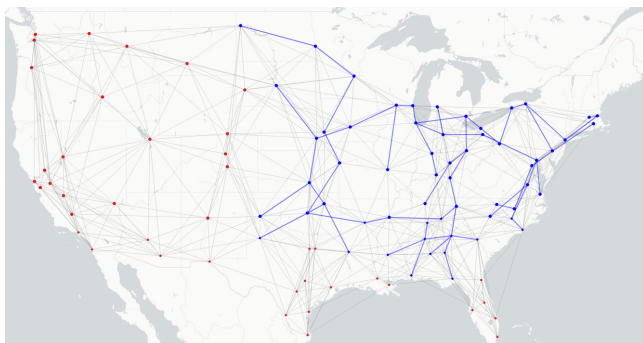
$$\lambda_1 = -10^{-5}, t_{max} = \infty$$



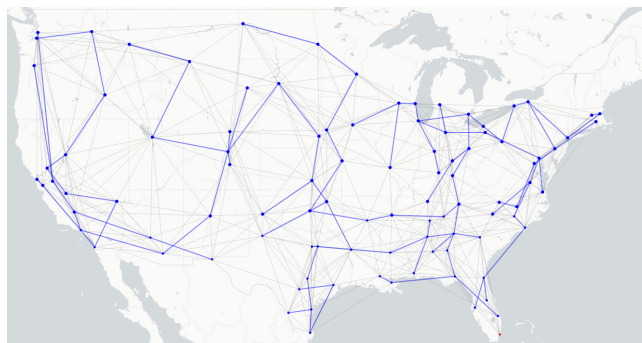
$$\lambda_1 = -5 * 10^{-5}, t_{max} = 500$$



$$\lambda_1 = -2 * 10^{-5}, t_{max} = 500$$



$$\lambda_1 = -2 * 10^{-5}, t_{max} = 8000$$



$$\lambda_1 = -2 * 10^{-5}, t_{max} = \infty$$

Figure 6: Simulated cascades in the random and small-world networks under different sets of λ_1 and t_{max}

the small-world network are highly connected with a small average geographic distance. It is also intuitive to observe that the contagion spreads like a **chain via the connected cliques** in the small-world network but more **divergent** over space in the random network. When the cutoff diffusion time is Inf , still the node 40 (i.e., Raleigh) in the random network and the node 41 (i.e., Miami) in the small-world network haven't been infected. This is because some nodes only have outward connections. Contagions cannot spread to these nodes unless they are the starting infected node.

7 DISCUSSION

First novel aspect of the STAND algorithm is that the geographic distance is implemented into the hazard rate function, **integrated over the time** to quantify its influence on the diffusion probability in Equation (9). By varying its coefficient λ_1 , we can model different real-world diffusion phenomena, where the geographic aspects have different effects.

The second novel aspect is that it can be applied to any kind of network structures that are observed in the real-world diffusion phenomena. The network structure is regarded as the input, and the hazard rate and diffusion probability are calculated separately for each single edge dependent on the distance and the infection time difference between two nodes that this edge connects. This is more flexible and accurate than previous research that assume a uniform global diffusion rate over the network.

The third novel aspect of the STAND algorithm is that the node properties and spatial dependencies (e.g., geographic distance) are considered to be vectors of explanatory variables X and Y in the definition of the hazard rate in Equation (6). Our STAND algorithm allows different node properties or spatial dependencies for different node-pairs within the same network since the diffusion probability is calculated separately for each edge.

8 CONCLUSIONS AND FUTURE WORK

This research presents a network diffusion algorithm, called STAND, based on the framework of probabilistic survival modeling, to simulate the diffusion cascades and infer the underlying spatiotemporal network over which various contagions (e.g., political policies, social opinions, news, diseases, etc.) spread over both space and time. It can estimate the diffusion speed and the **individual-level** infection likelihood given any network structures as input. It can help discover the diffusion pathways and also predict new occurrences of infections over space and time as well as trace back to detect where and when the diffusion began.

Real-world network diffusion phenomena are usually more complex than the simulated scenarios we have described here. The complexities can be summarized as follows: 1) the real-world networks over which contagions spread are a mixture of different levels of randomness and clustering rather than the simple random network or small-world network structure; and 2) the network nodes are heterogeneous and their properties have high influence on the diffusion process. We will address these complexities as the next step in our research. Future work will collect large-scale real-world spatiotemporal diffusion dataset and take various geographic characteristics into consideration, such as the local population or geographic scale.

Also, a more complex interactive visualization tool and an R package that integrates the STAND algorithm should be developed in the future to model the synthetic spatial network and simulated diffusion cascades as an aid to understanding high-dimensional and complex results.

DATA AVAILABILITY

We have published our codes and data at the Spatiotemporal Network Diffusion Codes and Data repository on the ScholarSphere. This repository collection includes: 1) 100 geographic locations we used for modeling the network nodes; 2) R codes for simulating both the random network and small-world network as well as the diffusion cascades spreading over these two network structures, using the proposed STAND algorithm; and 3) CSV files of simulated diffusion cascades under different sets of parameters λ_1 and t_{max} .

ACKNOWLEDGMENTS

We would also like to thank to Prof. Guido Cervone who provided many suggestions and comments about this research.

REFERENCES

- [1] Zack W Almquist. 2018. Large-scale spatial network models: An application to modeling information diffusion through the homeless population of San Francisco. *Environment and Planning B: Urban Analytics and City Science* (2018).
- [2] Christopher L Barrett, Keith R Bisset, Stephen G Eubank, Xizhou Feng, and Madhav V Marathe. 2008. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC'08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*. IEEE, 1–12.
- [3] Paul Erdos and Alfréd Rényi. 1960. On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences* 5, 1 (1960), 17–60.
- [4] KM Ariful Kabir and Jun Tanimoto. 2019. Analysis of epidemic outbreaks in two-layer networks with different structures for information spreading and disease diffusion. *Communications in Nonlinear Science and Numerical Simulation* (2019).
- [5] Elmar Kiesling, Markus Günther, Christian Stummer, and Lea M. Wakolbinger. 2012. Agent-based simulation of innovation diffusion: a review. *Central European Journal of Operations Research* (2012).
- [6] Jure Leskovec and Eric Horvitz. 2014. Geospatial structure of a planetary-scale social network. *IEEE Transactions on Computational Social Systems* 1, 3 (2014), 156–163.
- [7] David Liben-Nowell and Jon Kleinberg. 2008. Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences* 105, 12 (2008).
- [8] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences* 102, 33 (2005).
- [9] Vijay Mahajan, Eitan Muller, and Yoram Wind. 2000. *New-product diffusion models*. Vol. 11. Springer Science & Business Media.
- [10] Hazhir Rahmandad and John Sterman. 2008. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. *Management Science* 54, 5 (2008), 998–1014.
- [11] Manuel Gomez Rodriguez, David Balduzzi, and Bernhard Schölkopf. 2011. Uncovering the temporal dynamics of diffusion networks. *The 28th International Conference on Machine Learning* (2011).
- [12] Benjamin Taylor and Barry Rowlingson. 2017. spatSurv: An R package for Bayesian inference with spatial survival models. *Journal of Statistical Software* 77, 4 (2017), 1–32.
- [13] Zheyue Wang, Xinyue Ye, and Ming-Hsiang Tsou. 2016. Spatial, temporal, and content analysis of Twitter for wildfire hazards. *Natural Hazards* 83, 1 (2016), 523–540.
- [14] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of small-world networks. *Nature* 393, 6684 (1998).