Modeling the spread of infection in public transit networks: A decision-support tool for outbreak planning and control

Andras Bota
School of Civil and Environmental Engineering, The University of New South Wales,
Sydney, NSW, 2052, Australia, a.bota@unsw.edu.au

Lauren M. Gardner (corresponding author)
School of Civil and Environmental Engineering, The University of New South Wales,
Sydney, NSW, 2052, Australia, l.gardner@unsw.edu.au

Alireza Khani
Department of Civil, Environmental and Geo- Engineering
University of Minnesota Twin Cities
500 Pillsbury Drive SE, Minneapolis, MN 55455, USA, akhani@umn.edu

Word Count

6,250 + 3 Figures and 3 Tables = 7,750 total words

Submitted for Presentation and Publication
at the 96th Annual Meeting of the Transportation Research Board

Abstract

Identifying components of the public transit system most likely to exacerbate disease spread is critical for public health authorities to be able to plan for epidemics and control their spread. In this work we propose a method to detect such components in a transit network using a three-stage approach. We first use results from a transit simulation model to generate a contact network that is representative of the potential physical encounters between passengers. We then design and run a variety of epidemic scenarios atop this network which vary in their set of initially infected individuals, and properties of the disease. The expected spread (and variance) of the outbreak in terms of the infected passenger set are compared across scenarios. The infected passenger sets are then used to identify the vehicle trips most likely to transport infected passengers during an epidemic, which represent optimal locations to implement vehicle surveillance. Results from a case study using the public transit network from Austin, TX are presented. We show that the infection scenarios vary substantially in the pattern of spread at the individual level, but the set of transit vehicle trips at highest risk of infection is robust to the initial conditions of the outbreak.

1. INTRODUCTION

In many parts of the world a large proportion of the population lives and commutes in increasingly dense conditions, ideal for rapid disease transmission. Moreover, a major challenge is faced when planning for and controlling infectious disease outbreaks due to the stochastic nature of the spreading process (i.e., contact between an infectious and susceptible person may or may not result in a new infection) combined with the lack of information traditionally available on individual's daily contact patterns. For these reasons it is difficult to predict the impact that a new disease might have on a region, and similarly, design the best control strategies for public health authorities to implement during an outbreak.

Of particular interest in this study is the role of the public transit system in the spread of an infectious disease, which can act as a potential catalyst in the transmission process within metropolitan regions. We explore the risk posed by a regional public transit system in the event of an epidemic, with the goal of identifying potential vulnerabilities or critical components in the transit system (*e.g.*, super spreading vehicle-trips), which can be prioritized for monitoring and control during an emerging outbreak (1, 2).

We propose a network structure to represent the passenger contacts and movements within the public transit system. The network generated is used to model outbreak behavior within the transit system. A range of outbreak scenarios are simulated, and the most likely sets of infected passengers and vehicle trips are evaluated. We present results for a case study which uses transit simulation model results from Austin, TX.

2. BACKGROUND

The spread of contact-based infectious diseases can be modelled as a dynamic process atop social contact networks (3-7). Various agent based simulation models have been developed and implemented to replicate possible spreading scenarios, predict average spreading behavior, and analyze various intervention strategies for a given network and disease (8-22). Similar models have also been used to explain the propagation of information, ideas and opinions between individuals (23, 24).

To accurately predict disease spread, models must exploit characteristics of the population and the disease itself; and thus require large amounts of data, and significant computational resources. However, computational capabilities are exponentially increasing, and developments in data collection are increasingly providing a means for accurate mappings between known individuals. In particular, spatial analysis of transport and communication networks which exploit these data sources are a growing area of research (25-38). The ongoing development of activity-based travel models, which examine why, where and when various activities are engaged in by individuals (39-43), as well as innovations in pedestrian modeling (44) also present additional promising alternatives to generate social-contact networks in the future. These efforts have recently been expanded to include social network modeling, specifically the ability to reproduce spatial structure and interaction between individuals for large-scale social networks (42, 45-47).

Of particular applicability to this study are the recent advances in public transit modeling, which are able to provide detailed contact patterns, including temporal patterns (e.g. bus travel time) and spatial patterns (a function of the vehicle size and passenger volume) (48, 49). While these methods can potentially allow accurate mappings between known individuals, the data collection and processing required to recreate a real-world contact

network is expensive, computationally costly, and time intensive, among other challenges faced such as privacy issues (50-55). As such, it is critical to develop methods which can exploit this data as it becomes available, which is the contribution of this study.

Examples of such approaches have been recently addressed in the literature. Wesolowski et al. (56) highlighted the value of mobile network data for modelling human mobility patterns in real time, the results of which can be used in designing surveillance and containment strategies during an outbreak. The authors constructed intra and international mobility patterns in the set of African countries affected by and surrounding the Ebola outbreak. Their work exemplifies the role of increasingly available real-time connectivity data in public health response. Similarly, Sun et al (57) generated a physical encounter network using daily transit data from Singapore smart cards, and then used the contact network generated to compare different individual-based sensor schemes for early detection of simulated outbreaks (58). Their work utilized a "friends as sensors" approach, similar to that in (59-63), applied atop a more complete physical contact network.

Our study takes a step further in exploiting a public transport data set for planning city level disease mitigation and control efforts. Whereas Sun et al. (58) attempted to identify the individuals in the contact network that would serve as the best predictors of an outbreak, this work seeks to identify the *components of the transit system* that are most vulnerable (*i.e.*, most likely to have infected riders) during an outbreak (or attack). This analysis provides a means to understand what role the different components of the transit system play in the spreading process, and better predict how an outbreak might evolve over space (particular bus routes) and time.

The contact network of passengers is generated using data from Capital Metropolitan Transportation Authority in Austin, TX. The nodes of the network represent passengers, and they are connected if they ride on the same vehicle (bus or train) during their travels. The links are undirected with two attributes attached to them: the duration and the start time of the contact. A range of outbreak scenarios, which vary based on the number of location of initially infected individuals and properties of the disease, are then modelled atop this contact network, and the resulting infection patterns (*i.e.,* set of individuals and bus routes most likely to be infected) are compared.

We divide the analysis of the urban transportation network into two stages. The first stage quantifies network-based statistics on the contact network described above, including contact duration patterns, centralities and group structure. The second stage of the analysis focuses on modelling epidemics. A traditional discrete compartmental model (SIR) is implemented to simulate the spread of infections between individuals based on contact patterns. The sensitivity analysis reveals the robustness of the system to the initial conditions of a particular outbreak (*i.e.,* random versus targeted bus routes, and level of infectiousness of the disease).

The remainder of this manuscript includes a description of the data used in the study, the methodology implemented, followed by numerical results and conclusions.

3. DATA

In order to model passenger movement in the transit network, we used a schedule-based transit assignment model. The key point in such models is the time-dependent format of the network, in order to represent service variability during the day as well as modeling transit vehicle trips, as opposed to routes. This is particularly important because first, passengers'

route choice is more likely affected by service schedule and second, simulating their movement in the network requires knowing the particular vehicle they board. The transit assignment model, FAST-TrIPs (64) uses General Transit Feed Specifications (GTFS) data for the Austin, TX region in 2015 to build the schedule-based network. Passenger movement is then modeled by a previously estimated route choice model (65), in which parameters such as in-vehicle time, waiting time, walking time (for access, egress and transfer), and transfer penalty are incorporated. Finally, passengers' movement is simulated in the schedule-based network such that passengers are loaded to the network according to their chosen path. During simulation, passengers are modeled as agents, were created and loaded to a FIFO queue at their first boarding stop, boarded the assigned vehicle-trip upon its arrival, alighted at the alighting stop, and walked to either the next boarding stop or the destination. This process is modeled by an event-based simulation in which vehicle arrivals and departures are events as well as passengers' arrival at the stop (or destination), boarding, and alighting. Clearly, the model captures the interaction between passengers on the vehicles and at the boarding stops.

The transit assignment model was developed for Austin TX region. It involves a network of 75 routes (one commuter rail, two arterial BRT, 7 express bus routes, and 54 local bus routes, and 11 university shuttle routes) and 4,859 vehicle trips serving 2,653 stops in a weekday 24-hour period. On the demand side of the model, we used the transit origin-destination trip matrix from the 2015 travel demand model developed by the metropolitan planning organization in the region. In order to prepare the data for the assignment model, synthesized passengers were generated using the OD data and an aggregate departure time profile obtained from passenger count data. The assignment model was run and 88,570 passengers were assigned to the network, resulting in 111,640 unlinked passenger trips.

The transit assignment model generates various detailed outputs including the load profile of vehicles and passenger trajectories. In the load profile output, for each vehicle trip, the number of boarding and alighting passengers at each stop along with the time stamp and estimated vehicle dwell time is reported. Therefore, one can know how many passengers are on-board a vehicle in each segment of the route. The passengers' trajectory output, for each passenger, contains every activity during the course of the passenger trip along with stop ID, vehicle ID, and time stamp. Therefore, one can map the passenger activities to the schedule-based network and capture the interaction between passengers at each point in the space-time diagram.

4. NETWORK STRUCTURE

The collected transit data naturally defines a network structure. Passenger trajectories from the assignment model output are processed to create a contact network. Each passenger is considered as a node in the contact network, and links connect any two passengers that are on the same vehicle for a positive time period (i.e. in a segment of the corresponding transit route). The complete network is comprised of 88,570 nodes and 2,260,373 links.

For the purposes of our analysis, the time dimension of the network is defined on each link by two components: a *time stamp*, corresponding to the time when the second passenger gets on the vehicle while the first passenger is already on-board, and the *contact duration*, which indicates the temporal overlap between two passengers, *i.e.*, the time until one of the passengers gets off the vehicle. In this study we focus on the contact patterns between passengers in the morning peak period between 5 AM and 10 AM, therefore we only consider

the subgraph induced by the set of links where the time stamp indicating the start of the contact is within this period. We denote this network as the *contact network*. The contact network contains 31,176 nodes (passengers) and 1,006,459 links (contacts).

The id of the *vehicle trip* is an additional link attribute representative of where the contact takes place. The vehicle trip refers to a specific route with a specific departure time, and is specific to a single vehicle. Thus, the contact network includes a set of vehicle trips assigned to each passenger, and the set of passengers on each vehicle trip. We denote the passengers travelling on multiple trips as *transfer passengers*, while the passengers utilizing only a single vehicle trip as *bulk passengers*.

5. NETWORK STATISTICS

Before discussing epidemic spreading on the network, we examine the structure of the generated network in detail.

5.1 Contact network

We mentioned before that the contact network represents the morning peak behavior of passengers on an urban transportation network on a single workday. For each contact the start time and the duration is available, as well as the id of the vehicle where the contact takes place.

The density plot of the contact start times and durations are illustrated in Figure 1/a and 1/b. The start time distribution peaks at 7 AM and remains high until 9AM, which is consistent with a morning weekday commute. The majority of the contacts are in the magnitude of several minutes. The degree distribution of the network is shown in Figure 2/c. It follows a skewed power law, where the average number of contact per person is 65 while the maximum is 348. More than three quarters of the passengers do not transfer: they are bulk passengers, more than 20% of the travelers' only change vehicles once, and less than 2% change vehicles two or more times. While the number of transfer passengers is low, their role is critical; they represent the set of individuals that connect different parts of the network together allowing the transmission of diseases.

The clique size distribution can be seen in Figure 2/d. Since every passenger travelling on a vehicle at the same time is connected and the number of transfer passengers is relatively low, the network can be considered as a union of large overlapping cliques. In fact, if we examine the sizes of maximal cliques we can find that there are only 137 2-cliques (single edges) and around 300 triangles; the rest of the network is covered by cliques larger than this. If we group the nodes together according to vehicle trips, the resulting sets will be very dense and the overlaps between them (the transfer passengers) will be small, further emphasizing this structure.

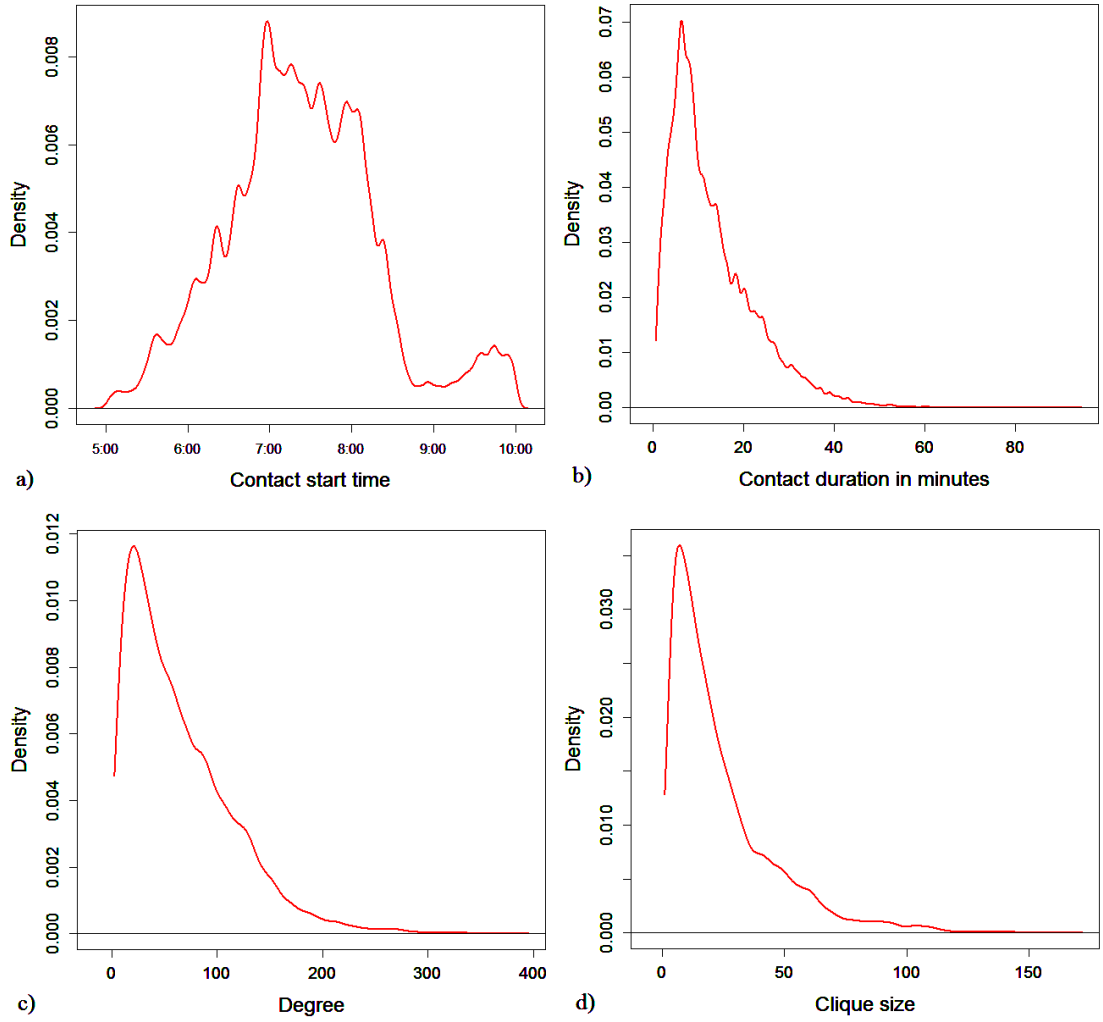

Figure 1. The distribution of (a) contact start times, (b) contact durations, (c) the degree distribution and (d) clique size distributions of the contact network.

Since the network is too large to be appropriately visualized, we illustrate this phenomenon on the two subgraphs of Figure 2. Figure 2/a illustrates the passenger with the most contacts, which is the black node in the middle, and its direct neighborhood. The nodes are colored according to vehicle trips, and darker edges indicate longer contact durations. It is clear that nodes are easily grouped according to vehicle trips. In reality these nodes were present on multiple vehicle trips and the color only indicates the first vehicle trip with the central passenger. This graph again illustrates the low number of transfer passengers.

Figure 2/b illustrates the contact network of the vehicle trip with the highest ridership. Here, the nodes are colored according to the contact start time, with lighter colors corresponding to earlier connections. The links are still shaded according to contact duration. This figure depicts how contacts on a single trip change in time. It reveals that the passengers that meet earlier in the morning tend to travel together for longer durations. It should be noted that not all of the passengers are on the vehicle at the same time. This figure also again highlights that the contact network on a single vehicle trip is very dense, and a union of large cliques.

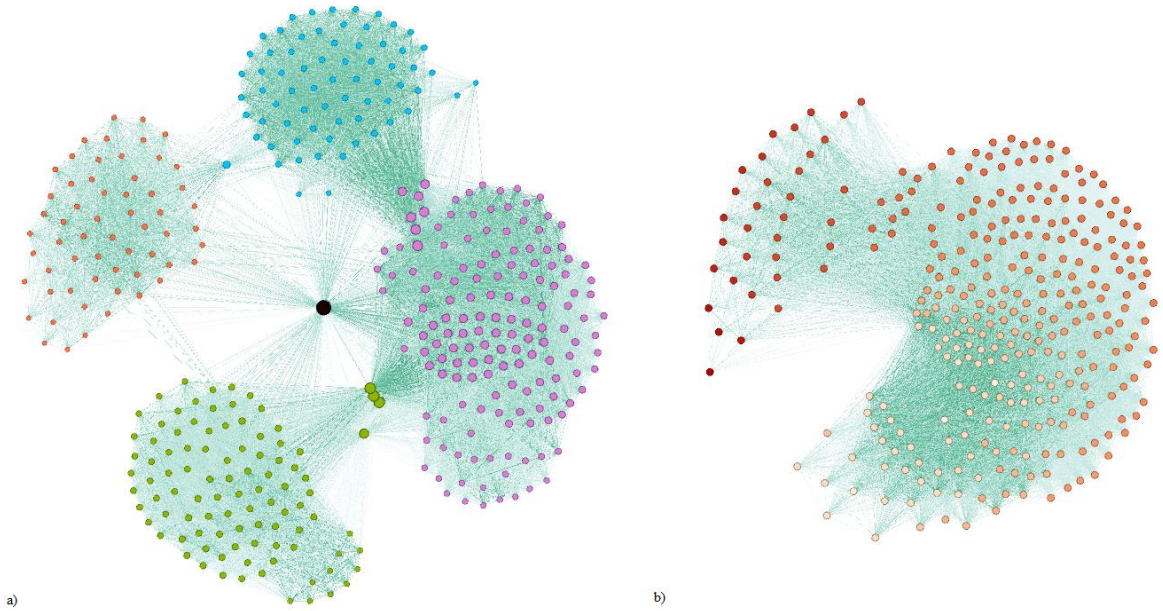

Figure 2 Two subgraphs of the contact network. 2/a The passenger with the most contacts is represented by the black node in the middle, connected to all contacts. The colors are indicative of the vehicle trips the passengers first met on, with darker edges representing longer contact durations. 2/b illustrates the contact network of the most frequented vehicle trip where colors indicate contact start times.

6. MODELING THE SPREAD OF INFECTION

The focus for the remainder of this paper is the analysis of epidemic spreading scenarios within the public transit network. For this purpose, we simulate epidemic spreading on the contact network, which is assumed to represent passenger travel patterns on a typical morning workday commute. A five-day observation period is considered to represent a full work week, and the contact network is assumed to remain constant across days. We consider multiple infection scenarios differing in both the size and spatial distribution of the initially infected individuals, and the level of infectiousness of the disease. The network-based infection model used to model the spread of disease between the individuals described in the following section is.

6.1 Infection model

A discrete compartmental susceptible-infected-removed (SIR) model is used to model an epidemic on the network, and is defined as follows. The SIR infection process takes place on a network defined as $G(V, E)$, where $V(G)$ denotes the node set of G, and $E(G)$ denotes the edge set of G. The process can be defined on directed and undirected networks, but in this study we only consider the latter case. Real values denoted as edge infection probabilities are required on the edges of the network, these are defined as $w_e \in [0,1]$ for all $e \in E(G)$. An edge infection probability w_e on an edge $e(u, v)$ denotes the probability that if node u is infected, the infection spreads to node v, or if v is infected, the infection spreads to u.

During the SIR infection process all nodes are required to adopt one of the three available states. A node can be susceptible (S), meaning it is not infected and can become infected, infected (I), meaning it is infected and may spread the disease to neighboring nodes, or removed (R) which means it was previously infected and has since recovered and is no longer

able to spread infection. A discrete time period is attached to the infected state denoted as τ_i indicating the amount of time a node spends in this state. The non-empty set of *initially infected* nodes will be denoted as $A_0 \subseteq V(G)$; these nodes are the *sources* of the infection. We define an *infection scenario* as the triple $S : (G, W, A_0)$, where G is a graph, $W : E(G) \mapsto [0,1]$ is a surjective assignment of edge infection probabilities to the edges of the graph, and A_0 denotes the set of initially infected nodes. The infection scenario defines the input of an infection process.

The infection process itself is iterative and takes place in discrete time steps, according to the following logic. Let $A_i \subseteq V(G)$ denote the set of infected nodes in iteration i. Each infected node $u \in A_i$ tries to infect its susceptible neighbors $v \in V(G) \setminus \bigcup_{0 \le j \le i} A_j$ according to w_e. If the attempt is successful v will be in an infected state starting from iteration $i + 1$. If the attempt is unsuccessful, node u it may make additional attempts to infect v depending on i and τ_i. If more than one node is trying to infect v in the same iteration, the attempts are made in an arbitrary order independent of each other. Finally, all nodes in an infected state change their state to removed if the time they became infected t_u is $t_u = i - \tau_i$.

The application of the SIR infection model to the contact network in this study can be interpreted as follows: Each iteration corresponds to a single day. If a person gets infected in iteration i then he is infectious for τ_i days, starting from iteration $i + 1$, indicating a latency period of a single day[1]. In this study we assume an infectious period (τ_i) of five days. Results from the process provide the state of each individual (S, I or R) and each iteration. Due to the five-day observation period, any infected individual remains infectious throughout the entire simulation, i.e., there will be no recovered individuals within the timeframe considered.

Due to the stochastic nature of the model, we seek to estimate the likelihood of infection for each node at each iteration. In this study for each outbreak scenario considered we run the infection process k times and count the frequency of infection for each node at any given iteration to estimate the probability that a node is in an infectious state at any iteration. For all results in this paper $k = 100{,}000$ was used. This approach is similar to that used in (66, 67).

6.2 Experiment Design

In order to analyze the effect of an epidemic on the public transportation system we consider a diverse set of infection scenarios on the contact network. The scenarios represent a high variety of initial outbreak conditions, which can be divided into two strategies according to the definition of the infection scenario in the previous section. The edge infection probability assignment represents the level of infectiousness of a particular disease, while the set of initially infected nodes represents the source(s) of the outbreak. The set of initially infected nodes can be further divided into two factors: the size of the initial set and the spatial distribution of it.

Following the assumption of (57) we define the transmission probabilities as a linear function of the contact duration between individuals: $p_{u,v} = d_{u,v} * \beta$, where $p_{u,v}$ is the transmission probability between nodes u and v, $d_{u,v}$ is the contact duration between u and v in minutes, and β is a constant. We consider four different values for β which we believe represent a

[1] While in the infection model, this translates into a latency period of zero iterations.

wide range of outbreaks. $\beta = 0.001$ which puts the edge infection values between 0 and 0.1, $\beta = 0.003$ which puts the transmission probabilities between 0 and 0.3, $\beta = 0.005$ with the bounds 0 and 0.5 and $\beta = 0.007$ with 0 and 0.7.

The size of the initially infected node set A_0 will be 10, 50 and 100 representing a small, medium and large number of initially infected individuals. These individuals are distributed according to two different strategies, similar to the works of (68). The first strategy considers the top 10, 50 and 100 most central[2] individuals corresponding to a "targeted" strategy, while the second strategy selects the same amount of nodes randomly. In the random selection strategy ten samples were drawn independently from $V(G)$ for each scenario[3], and the test results were averaged.

6.3 Limitations of the study

Before presenting the results certain limitations of this study should be noted. The first set of limitations is due to the available data. The passenger demand used in the assignment model (used to generate the contact network used in this study) was estimated using a traditional *trip-based* travel forecasting model; thus individuals could not be tracked between their morning and evening commute due to the lack of information about their daily tour. For this reason, only the morning peak period (rather than the entire day) was used in this study. Secondly, the contact network is generated using a single morning of data, but used to model the spread of an outbreak over a 5-day period. This implies the assumption that morning commute patterns remain constant day-to-day over a 5-day work week. While individual's travel patterns can vary day-to-day, a recent study (57) which used travel smart card data to generate an in-vehicle social encounter network on public buses using a full week of travel data found that physical encounters display reproducible temporal patterns. The finding that repeated encounters are regular and identical, and rooted in daily behavioral regularity supports the assumption used in this work. While the restriction to morning peak travel does limit the applicability of our results (we underestimate the spread of disease), the proposed methodology, which is the main contribution of this work, can be directly applied to a more complete contact data set. Planned future research will utilize a 24 hour (or longer) travel period that tracks the individual users over the entire period.

Implicit assumptions also result from the use of the SIR model, which restrict this study to the family of infectious diseases that are transmitted from an infected to susceptible individual via direct or close contact. This category includes various strands of the flu, SARS and the common cold, among others and, diseases which have an infectious period greater than five days. These assumptions could easily be relaxed without altering the proposed methodology, and a more complex compartmental model, *e.g.*, SEIR, could be substituted.

Finally, this study ignores all contacts made outside of transit movements, which in reality would significantly increase the size of the outbreak within a region (while also resulting in more infected transit riders). However, predicting the size of the outbreak is not the intended goal of this study. Instead, the contribution is the proposed methodology to identify the components of the public transit system that play a critical role during the early stages of an outbreak. The identified transit-trips can be prioritized for surveillance monitoring and

[2] Degree centrality was computed and used to identify the set of nodes.
[3] For each size configuration (10, 50, 100) ten different samples were taken.

possible mitigation and control efforts by transit and public health authorities during an outbreak.

7. RESULTS

We examine the results of the infection simulations in this section. Our key point of interest is the behavior of the outbreak as a function of the initial conditions. Apart from the fraction of infected individuals we are most interested in the components of the transit system that are at highest risk of infection. Due to space restrictions we present all results as of day 5; however analogous results for any stage of the outbreak are available.

7.1 Comparison in initial infection scenarios

Table 1 reveals the fraction of infected individuals in the contact network (on day 5) for each infection scenario considered. The rows correspond to the transmission probabilities, while the columns indicate the size of the initially infected set and the selection strategy. As expected, the prevalence of the outbreak increases with number of initially infected individuals, *i.e.*, size of A_0 , and the level of infectiousness of the disease, β. Results are also provided for the targeted and random strategies, which only vary based on the spatial distribution of the initial set of infected individuals.

Table 1. Fraction of infected individuals on the contact network on day 5. The rows denominate the constants used to compute the edge infection probabilities, while the columns indicate the size of the initially infected set and the selection strategy. The values in parenthesis for the random strategy indicate standard deviation.

	Targeted strategy			Random strategy		
	$\lvert A_0\rvert = 10$	$\lvert A_0\rvert = 50$	$\lvert A_0\rvert = 100$	$\lvert A_0\rvert = 10$	$\lvert A_0\rvert = 50$	$\lvert A_0\rvert = 100$
$\beta = 0.001$	0.0262	0.0612	0.1015	0.0068(58%)	0.0311(22%)	0.0631(12%)
$\beta = 0.003$	0.1171	0.2177	0.2921	0.0442(47%)	0.1693(15%)	0.2826(8%)
$\beta = 0.005$	0.2075	0.3381	0.4183	0.0993(45%)	0.3131(11%)	0.4512(5%)
$\beta = 0.007$	0.2847	0.4299	0.5051	0.1623(33%)	0.4297(8%)	0.5647(4%)

The results in Table 1 generally confirm previous observations (68) that a targeted strategy has a much greater impact on the network, than a random one. There is some difference in the finer details however. To illustrate this, we classify the behavior of the infection process into three categories. We assign the infection scenarios to the above categories as follows. The pairs indicate $(\beta, |A_0|)$.

I. $(0.001,10), (0.001,50), (0.001,100), (0.003,10), (0.005,10), (0.007,10)$.
II. $(0.003,50), (0.003,100), (0.005,50), (0.007,50)$.
III. $(0.005,100), (0.007,100)$.

In the first category (I) a targeted strategy has greater impact than the random one. This is true if the size of the initial outbreak or the network transitivity (edge infection values) is small, and also corresponds with the observations of (68). For example, if $\beta = 0.001, |A_0| = 10$ the difference is almost a multiplier of 4. For scenarios with high transitivity $\beta = 0.005, |A_0| = 10$ the observation is still true albeit of a lesser magnitude. This is because with a small number of sources or low levels of infectiousness, the set of the initially

infection individuals is critical. In other words, a small number of well-connected individuals chosen as source infections have a much more substantial impact on the system compared with the same number of randomly selected individuals.

The second category (II) corresponds to a broad range of configurations where there is little difference between targeted and random strategies. In this case the network transitivity and the size of the initial infections are such that the selection strategy for the source infections matters less; the infection spreads around the network to a similar number of individuals. On the contact network, this occurs when the prevalence of infection is moderately high.

In the third category the random infection strategy has greater impact than the targeted one. The explanation behind this somewhat counter-intuitive, and may be due to assortativity. In assortative networks nodes tend to be connected to other nodes with similar degrees. This is true for the nodes with high degrees as well, grouping them together. In contrast, random nodes are selected randomly so they are spread out across the network. Since the network transitivity is high and there are a considerable number of initial infections, the randomly selected (and more spread out) individuals may be able to infect more individuals than the well-connected (but grouped) targeted individuals.

The categorization becomes clearer if we look at the standard deviation values for the random selection strategy indicated by the values in parenthesis in Table 1. These represent how the expected size of the infection varies across the ten samples for each $|A_0|$ of a random scenario. High deviation values in the first category show that the selection of the initially infected set is critical and has a considerable effect on the outcome. For the second and third categories, this becomes less true, to the point, that it does not matter how the sources of the infection are chosen, the network "takes care" of the spreading process.

We examined the expected number of infected passengers on each individual day and found that in all infection scenarios the outbreak progresses in a near-linear fashion on subsequent days, with the slope of the ascent depending on the initial conditions. There is little difference between the categories.

7.2 Identification of critical components of the transit system

The previous section explored the evolution of the outbreak across individuals in the network. However, the main objective of this study is to identify the most critical bus trips responsible for furthering the spread of disease amongst passengers (and more generally within the region) during an outbreak. To identify these critical components of the transit system we compute and rank the vehicle trips according to their likelihood of transporting infected passengers.

The infection model defined in Section 6 provides the likelihood of infection for the nodes of the network, which represent individuals. Using the known assignment of the individuals to vehicle trips and the infection probability for each individual passenger, we compute the likelihood of infection on each vehicle trip by averaging the probability of infection for all passengers on a given vehicle trip at each iteration, *i.e.,* day. The vehicle trips are then ranked according to their corresponding probability of infection and the top 100 are considered. In this study 100 is representative of the top 10% of vehicle trips in the network we evaluate, but any threshold level could be chosen. The process is conducted for each infection scenario simulated.

We first compare the robustness of the highest risk vehicle trips *within* each scenario. The maximum number of times a vehicle trip appears in the top 100 ranking across the ten samples, and the number of vehicle trips with this maximum frequency are provided in the left hand side of Table 2, with the latter in parentheses. For example, 8(3) means that 3 different vehicle-trips appeared in the top 100 ranking in 8 of the ten samples (for a given random scenario). The right hand side of the table compares the overlap between each random strategy (using the top 100 most likely infected vehicle trips across all ten samples) and targeted strategy for the same initial outbreak conditions. It is clear for the more aggressive outbreaks (more contagious with a larger number of sources) the set of most likely infected vehicle trips appears in all ten samples, and is more likely to appear in the targeted strategy. Almost all outbreak scenarios share more than 50 vehicle trips across the respective random and targeted strategy, suggesting robustness in the set of riskiest vehicle trips. These results correspond to the standard deviations in Table 1, and further support the categorization introduced there.

Table 2. The maximum number of times a vehicle trip appears in the top 100 ranking across the ten samples, and the number of vehicle trips with this maximum frequency are provided in the left hand side for the random strategy. The right hand side represents the number of vehicle trips identified in both the random and targeted strategies.

	Random Strategy Frequencies			Random -Targeted Frequencies		
	$\|A_0\| = 10$	$\|A_0\| = 50$	$\|A_0\| = 100$	$\|A_0\| = 10$	$\|A_0\| = 50$	$\|A_0\| = 100$
$\beta = 0.001$	7(2)	8(3)	10(2)	42	57	55
$\beta = 0.003$	9(2)	10(4)	10(12)	55	69	69
$\beta = 0.005$	10(1)	10(10)	10(22)	56	69	73
$\beta = 0.007$	10(2)	10(14)	10(24)	60	68	74

In addition to identifying the list of highest risk routes for each scenario independently, we seek to identify a robust set of vehicle trips which are most likely to transport infected passengers across a range of outbreak scenarios (similar initial conditions). To identify this set we consider the different categories defined previously, and count the frequency of each vehicle-trip's occurrence in the top 100 vehicle trips across all scenarios in each category. The frequency list provides an aggregated ranking of the most likely infected vehicle trips across a range of scenarios. The frequency results for each category as well as the global set are illustrated in Table 3. There are a total of 132 scenarios considered[4], with Cat I, II and III corresponding to 60, 40 and 20 scenarios respectively. The vehicle trips which overlap across all 4 categories are highlighted in Table 3.

The global aggregation reveals that there are 12 particular vehicle trips which occur in over 100 of the 132 scenarios, implying this particular set of routes are highly likely to have infected travelers, irrespective of where the infection starts (random vs. targeted, $|A_0|$) or the level of infectiousness of the disease in question. Similarly, high occurrence of particular vehicle trips can also be seen in the other three categories. The first category contains 60

[4] Four β values, three $|A_0|$ settings, 10 random samples of A_0 and one targeted for each $|A_0|$.

infection scenarios and the most common (two) vehicle trips are present in 48 of them. Category II contains 40 scenarios and three trips contained by all of them. The last category contains 20 scenarios and 21 trips are present in all of them. This again confirms our previous results that the overlap is greater between infection scenarios in the second and third categories but even in the first one we are able to identify several vehicle trips where the infection is likely to be present. if we consider infection scenarios in category I, the overlap between them is surprisingly high. For example, in the scenario with $\beta = 0.001, |A_0| = 10$ ten passengers are selected randomly out of 31000 and still seven times out of ten, the same (two) vehicle trips carry the infection. This further confirms our observation; the outbreak has a tendency to reach the same vehicle trips in the transportation network independent of the initial conditions.

Table 3. The global and categorical rankings of the vehicle trips where infection is most likely to be present. The numbers in the table indicate the ID of the route, the start time for the specific vehicle trip, and the frequency (number of scenarios) of the trip being infected. Trips that appear in all rankings are bolded in red.

Rank	Global (132)	Cat. I. (60)	Cat. II. (40)	Cat. III. (20)
1	**19 - 7:05 (120)**	**1 - 6:04 (48)**	**19 - 7:05 (40)**	661 - 7:38 (20)
2	**1 - 6:04 (119)**	**19 - 7:05 (48)**	**19 - 6:33 (40)**	**19 - 7:05 (20)**
3	**19 - 5:58 (116)**	**323 - 7:25 (46)**	**383 - 6:05 (40)**	**19 - 5:58 (20)**
4	**323 - 7:25 (114)**	**19 - 5:58 (45)**	**19 - 5:58 (39)**	**19 - 6:33 (20)**
5	**383 - 6:05 (110)**	30 - 7:10 (41)	**1 - 6:04 (39)**	20 - 6:42 (20)
6	**19 - 7:08 (109)**	**19 - 7:08 (40)**	**30 - 7:53 (39)**	**103 - 6:25 (20)**
7	**19 - 6:33 (108)**	333 - 7:15 (38)	**19 - 7:08 (38)**	**19 - 7:08 (20)**
8	30 - 7:10 (107)	**103 - 6:25 (38)**	661 - 7:58 (38)	**1 - 6:04 (20)**
9	**103 - 6:25 (105)**	**383 - 6:05 (38)**	661 - 7:38 (37)	**323 - 7:25 (20)**
10	**30 - 7:53 (104)**	983 - 7:30 (37)	**19 - 7:40 (37)**	**19 - 7:40 (20)**
11	983 - 7:30 (102)	970 - 7:25 (36)	142 - 6:30 (37)	**103 - 6:45 (20)**
12	**19 - 7:40 (101)**	663 - 8:02 (36)	1 - 6:52 (36)	19 - 6:15 (20)
13	661 - 7:38 (99)	**19 - 7:40 (36)**	30 - 7:10 (36)	19 - 5:21 (20)
14	**103 - 6:45 (99)**	**19 - 6:33 (36)**	**323 - 7:25 (36)**	142 - 6:30 (20)
15	142 - 6:30 (99)	**103 - 6:45 (35)**	**103 - 6:45 (36)**	**383 - 6:05 (20)**
16	661 - 7:58 (98)	935 - 6:45 (34)	1 - 6:26 (36)	171 - 6:43 (20)
17	1 - 6:26 (97)	**30 - 7:53 (34)**	323 - 6:50 (36)	1 - 6:26 (20)
18	970 - 7:25 (97)	661 - 7:58 (33)	**103 - 6:25 (35)**	323 - 6:50 (20)
19	1 - 6:52 (95)	1 - 6:52 (32)	983 - 7:30 (34)	383 - 5:35 (20)
20	19 - 6:15 (93)	21 - 7:20 (32)	970 - 7:25 (34)	**30 - 7:53 (20)**

These transit routes with the highest likelihood of infection are visualized in Figure 3. Each subfigure illustrates a category of the routes listed in Table 2 (i.e., the categories). The width of the lines in the figure shows the ridership of the route in each segment, and the color indicates the frequency of the vehicle trip (with a spectrum from blue to yellow to red, with red representing the highest number). It can be observed that a subset of routes is common among categories with slight differences. Notably, most of the routes serve the central area where downtown and the university campus is located. More specifically, routes 19 and 103 are present in all the categories with high frequency of being infected. Route 19 is a local

route serving residential areas where university students live and the university campus and downtown. Route 13 connects high density corridors to the downtown.

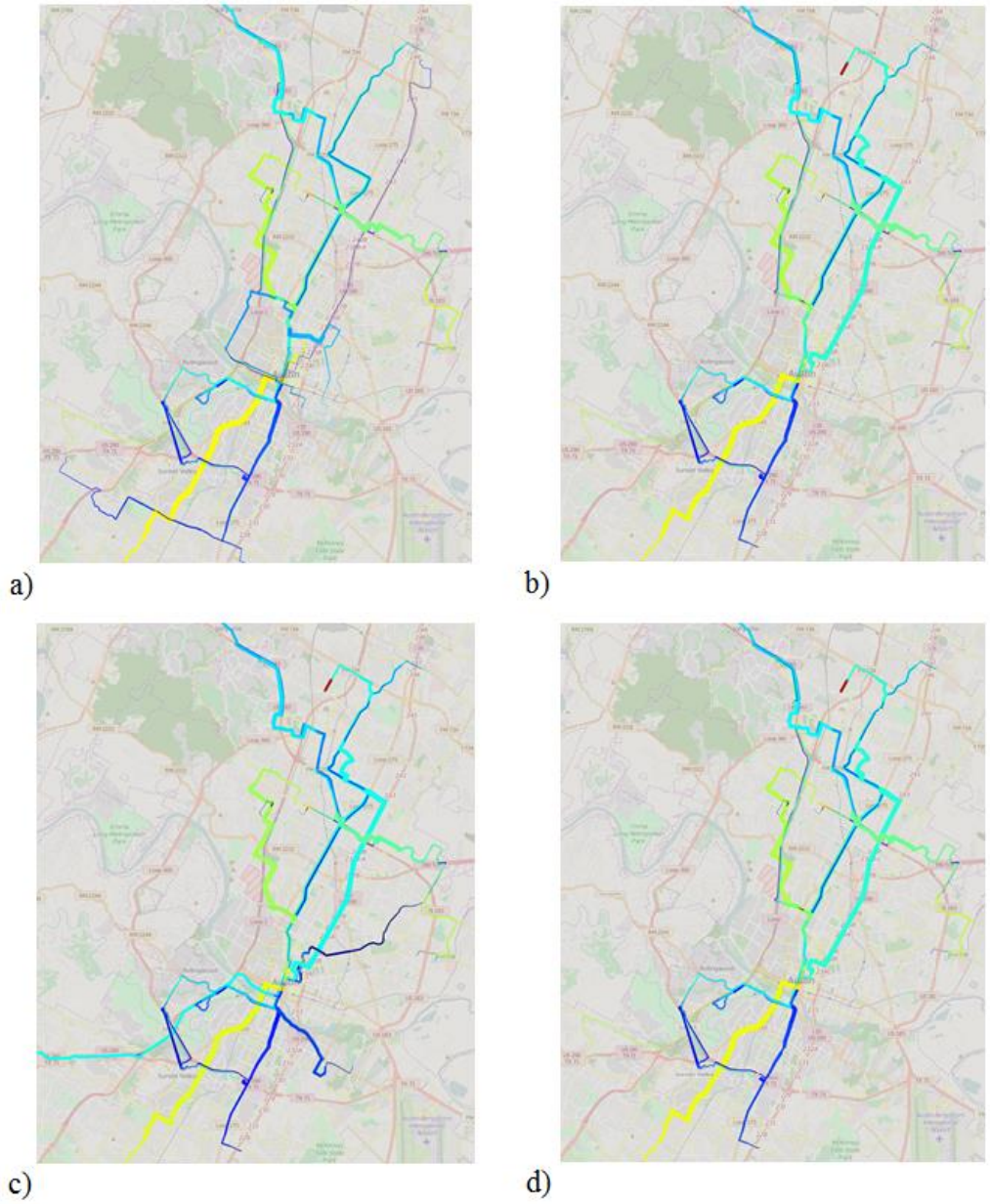

Figure 3. The transit routes with the highest frequency of being infected. Categories I – III can be seen on a – c), respectively, and d) depicts the top routes of the global ranking.

8. CONCLUSION AND FUTURE WORKS

We presented a novel method to identify the components of a public transit system which are most likely to transport infected passengers, and therefore play a vital role in the continued spread of infection during an outbreak. We ran a variety of epidemic scenarios atop the generated network which varied in their selection strategy of infection sources, number of the initially infected individuals, and levels of infectiousness. Based on the results from a case

study using the public transit network from Austin, TX, the infection scenarios can vary substantially in the pattern of spread at the individual level, but the set of transit vehicle trips at highest risk of infection is robust to the initial conditions of the outbreak. These vehicle-trips would be optimal locations to implement vehicle surveillance, for example in the forms of temperature gauges such as those recently installed on buses in Hong Kong. Identifying the critical spreading components of the transit system can aid public health authorities in the optimal allocation of surveillance resources, and is essential in mitigating future outbreaks. The application in this study has recognized limitations due to the available data, but the proposed method serves as a proof of concept, which can be directly applied to more complete travel data sets, and also extended to include additional sources of human contact outside of travel if a more inclusive contact data were available.

References

1. Galvani AP, May RM. Epidemiology: dimensions of superspreading. Nature. 2005;438:293-5.
2. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nat Phys 2010;6:888–93.
3. Murray J. Mathematical Biology, third ed. New York, NY: Springer; 2002.
4. Anderson RM, May RM. nfectious Diseases of Humans: Dynamics and Control. Oxford, UK: Oxford University Press; 1991.
5. Fajardo D, Gardner L. Inferring Contagion Patterns in Social Contact Networks with Limited Infection Data. Netw Spat Econ. 2013;13(4):399-426.
6. Gardner L, Fajardo D, Waller ST. Inferring Contagion Patterns in Social Contact Networks Using a Maximum Likelihood Approach. Natural Hazards Review. 2014;15(3).
7. Rey D, Gardner L, Waller ST. Finding Outbreak Trees in Networks with Limited Information. Networks and Spatial Economics. 2016;16(2):687–721.
8. Rvachev L, Longini I. A mathematical model for the global spread of influenza. Math Biosci 1985;75:3-22.
9. Epstein J, DAT. C, al. e. Toward a Containment Strategy for Smallpox Bioterror: An Individual-Based Computational Approach. Brookings Institute Press 2002;2004:55.
10. Eubank S. Modelling disease outbreaks in realistic urban social networks. Nature. 2004;429:180-4.
11. Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. Proc Natl Acad Sci USA. 2004;101(42):15124–9.
12. Dibble C, Feldman PG. The GeoGraph 3D Computational Laboratory: Network and Terrain Landscapes for RePast. Journal of Artificial Societies and Social Simulation 2004;7(1).
13. Cahill E, Crandall R, Rude L, Sullivan A. Space-time influenza model with demographic, mobility, and vaccine parameters. Proc 5th Annual Hawaii Internat Conf Math, Statist, and Related Fields. 2005.
14. Dunham JB. An agent-based spatially explicit epidemiological model in MASON. Journal of Artificial Societies and Social Simulation. 2005;9(1):3.
15. Meyers L, Pourbohloul B, Newman MEJ, Skowronski D, Brunham R. Network theory and SARS: Predicting outbreak diversity. Journal of Theoretical Biology 2005;232:71-81.
16. Small M, Tse CK. Small world and scale free model of transmission of SARS. International Journal of Bifurcations and Chaos Appl Sci Eng 2005;15(1745).
17. Carley KM, Fridsma DB, Casman E, Yahja A, Altman N, Chen LC, et al. BioWar: scalable agent-based model of bioattacks. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions. 2006;36(2):252-65.
18. Ferguson NM. Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature. 2005;437:209-14.
19. Germann TC, Kadau K, Longini IM, Macken CA. Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci. 2006;103(15):5935–40.
20. Ekici A, Keskinocak P, Swann JL. Pandemic influenza response. Winter Simulation Conference. 2008:1592-600.
21. Roche B, Drake J, Rohani P. An Agent-Based Model to study the epidemiological and evolutionary dynamics of Influenza viruses. BMC bioinformatics 2011;12(1):87.

554 22. D.T. H, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, J. W, et al. The
555 construction and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth
556 outbreak. Proc R Soc B. 2003;270:121-7.
557 23. Coleman J, Menzel H, Katz E. Medical Innovations: A Diffusion Study. New York,
558 NY: Bobbs Merrill; 1996.
559 24. Hasan S, Ukkusuri SV. A contagion model for understanding the propagation of
560 hurricane warning information. Transportation Research Part B 2011;45(10):1590-605.
561 25. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL. Uncovering
562 individual and collective human dynamics from mobile phone records. J Phys A: Math
563 Theor. 2008;41(224015):11.
564 26. Gastner M, Newman M. The spatial structure of networks. Eur Phys J B
565 2006;49(2):247–52.
566 27. Schintler L, Kulkarni R, Gorman S, Stough R. Using Raster-Based GIS and Graph
567 Theory to Analyze Complex Networks. Networks and Spatial Economics 2007;7(4):301–13.
568 28. Erath A, Löchl M, K. A. Graph-Theoretical Analysis of The Swiss Road And Railway
569 Networks Over Time. Networks and Spatial Economics 2009;9(3):379–400.
570 29. Wang P, Gonzlez MC, Hidalgo CA, Barabsi AL. Understanding the spreading
571 patterns of mobile phone viruses. Science. 2009;324:1071-6.
572 30. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility
573 patterns. Nature. 2008;453:779-82.
574 31. González MC, Lind PG, Herrmann HJ. System of mobile agents to model social
575 networks. Physical Review Letters. 2006;96(8):088702.
576 32. Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature.
577 2006;439:462-5.
578 33. Song C, Qu Z, Blumm N, Barabási AL. Limits of predictability in human mobility.
579 Science. 2010;327:1018-21.
580 34. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD. Unique in the Crowd: The
581 privacy bounds of human mobility. Sci Rep. 2013;3:1376.
582 35. Chen N, Gardner L, Rey D. A bi-level optimization model for the development of
583 real-time strategies to minimize epidemic spreading risk in air traffic networks.
584 Transportation Research Record: Journal of the Transportation Research Board. 2016;No.
585 2569.
586 36. Gardner L, Fajardo D, Waller ST. Inferring infection-spreading links in an air traffic
587 network. Transportation Research Record: Journal of the Transportation Research Board.
588 2012(2300):13-21.
589 37. Gardner L, Sarkar S. A global airport-based risk model for the spread of dengue
590 infection via the air transport network. PLoS One. 2013;8(8):e72129.
591 38. Gardner LM, Fajardo D, Waller ST, Wang O, Sarkar S. A predictive spatial model to
592 quantify the risk of air-travel-associated dengue importation into the United States and
593 europe. J Trop Med. 2012;2012:103679.
594 39. Lam WHK, Huang H. Combined Activity/Travel Choice Models: Time-Dependent
595 and Dynamic Versions. Networks and Spatial Economics. 2003;3(3):323–47.
596 40. Roorda MJ, Carrasco JA, Miller EJ. An Integrated Model of Vehicle Transactions,
597 Activity Scheduling and Mode Choice. Transportation Research Part B. 2009;43(2):217–29.
598 41. Ramadurai G, Ukkusuri S. Dynamic User Equilibrium Model for Combined Activity-
599 Travel Choices Using Activity-Travel Supernetwork Representation. Networks and Spatial
600 Economics. 2010;10(2):273–92.
601 42. Illenberger J, Nagel K, Flötteröd G. The Role of Spatial Interaction in Social
602 Networks. Networks and Spatial Economics. 2012;13(3):1-28.

603 43. Huerta R, Tsimring LS. Contact tracing and epidemics control in social networks.
604 Phys Rev E Stat Nonlin Soft Matter Phys. 2002;66:056115.
605 44. Hoogendoorn SP, Bovy PHL. Pedestrian Travel Behavior Modeling. Networks and
606 Spatial Economics. 2005;5(2):193–216.
607 45. Balcan D. Multiscale mobility networks and the spatial spreading of infectious
608 diseases. Proc Natl Acad Sci Usa. 2009;106:21484-9.
609 46. Salathé M. A high-resolution human contact network for infectious disease
610 transmission. Proc Natl Acad Sci Usa. 2010;107:22020-5.
611 47. Funk S, Salathé M, Jansen VAA. Modelling the influence of human behaviour on the
612 spread of infectious diseases: a review. J R Soc Interface. 2010;7:1247-56.
613 48. Nassir N, Khani A, Hickman M, Noh H. An Intermodal Optimal Multi-Destination
614 Tour Algorithm with Dynamic Travel Times. Transportation Research Record: Journal of the
615 Transportation Research Board. 2012;2283:57-66.
616 49. Pendyala R, Kondhuri K, Chiu Y-C, Hickman M, Noh H, Waddell P, et al. Integrated
617 Land Use-Transport Model System with Dynamic Time-Dependent Activity-Travel
618 Microsimulation. Transportation Research Record: Journal of the Transportation Research
619 Board. 2012;2203:19-27.
620 50. Bajardi P. Human mobility networks, travel restrictions, and the global spread of 2009
621 H1N1 pandemic. P Lo S One. 2011;6:e16591.
622 51. Christakis NA, Fowler JH. Social network sensors for early detection of contagious
623 outbreaks. P Lo S One. 2010;5:e12948.
624 52. Cattuto C. Dynamics of person-to-person interactions from distributed rfid sensor
625 networks. Plo S One. 2010;5:e11596.
626 53. Stehlé J. Simulation of an seir infectious disease model on the dynamic contact
627 network of conference attendees. Bmc Med. 2011;9:87.
628 54. Kuiken C, Thakallapalli R, Eskild A, de Ronde A. Genetic analysis reveals
629 epidemiologic patterns in the spread of human immunodeficiency virus. Am J Epidemiol.
630 2000;152:814-22.
631 55. Gilbert MT. The emergence of HIV/AIDS in the americas and beyond. Proc Natl
632 Acad Sci Usa. 2007;104:18566-70.
633 56. Wesolowski A, Buckee C, Bengtsson L, Wetter E, Lu X, Tatem A. Commentary:
634 containing the Ebola outbreak–the potential and challenge of mobile network data. PLOS
635 currents outbreaks. 2014.
636 57. Sun L, Axhausen KW, Lee DH, Huang X. Understanding metropolitan patterns of
637 daily encounters. Proc Natl Acad Sci Usa. 2013;110:13774-9.
638 58. Sun L, Axhausen KW, Lee D-H, Cebrian M. Efficient detection of contagious
639 outbreaks in massive metropolitan encounter networks. Scientific Reports. 2014;4:5099.
640 59. Ginsberg J. Detecting influenza epidemics using search engine query data. Nature.
641 2008;457:1012-4.
642 60. Chan EH. Global capacity for emerging infectious disease detection. Proc Natl Acad
643 Sci Usa. 2010;107:21701-6.
644 61. Garcia-Herranz M. Using friends as sensors to detect global-scale contagious
645 outbreaks. 2012.
646 62. Shaman J, Karspeck A. Forecasting seasonal outbreaks of influenza. Proc Natl Acad
647 Sci Usa. 2012;109:20425-30.
648 63. Salathé M. Influenza A (H7N9) and the importance of digital epidemiology. N Engl J
649 Med. 2013;369:401-4.
650 64. Khani A, Hickman M, Noh H. Trip-based path algorithms using the transit network
651 hierarchy. . Networks and Spatial Economics, 15(3), 635-653. 2015;15(3):635-53.

65. Khani A, Beduhn TJ, Duthie J, Boyles S, Jafari E. A Transit Route Choice Model for Application in Dynamic Transit Assignment. Innovations in Travel Modeling Conference; Baltimore, MD2014.
66. Bóta A, Krész M, Pluhár A. Approximations of the Generalized Cascade Model. Acta Cybern. 2013;21(1):37-51.
67. Kempe D, Kleinberg J, E. T. Maximizing the Spread of Influence though a Social Network. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM. 2003:137-46.
68. Albert R, Barabási A. Statistical mechanics of complex networks. Reviews of modern physics. 2002;74(1):47.