The Network Source Location Problem in the Context of Foodborne Disease Outbreaks

Abigail L. Horn and Hanno Friedrich

Abstract In today's globally interconnected food system, outbreaks of foodborne disease can spread widely and cause considerable impact on public health. Food distribution is a complex system that can be seen as a network of trade flows connecting supply chain actors. Identifying the source of an outbreak of foodborne disease distributed across this network can be solved by considering this network structure and the dimensions of information it contains. The literature on the network source identification problem has grown widely in recent years covering problems in many different contexts, from contagious disease infecting a human population, to computer viruses spreading through the Internet, to rumors or trends diffusing through a social network. Much of this work has focused on studying this problem in analytically tractable frameworks, designing approaches to work on trees and extending to general network structures in an *ad hoc* manner. These simplified frameworks lack many features of real-world networks and problem contexts that can dramatically impact transmission dynamics, and therefore, backwards inference of the transmission process. Moreover, the features that distinguish foodborne disease in the context of source identification have not previously been studied or identified. In this article we identify these features, then provide a review of existing work on the network source identification problem, categorizing approaches according to these features. We conclude that much of the existing work cannot be implemented in the foodborne disease problem because it makes assumptions about the transmission process that are unrealistic in the context of food supply networks—that is, identifying the source of an epidemic *contagion* whereas foodborne contamination spreads through a transport network-mediated *diffusion* process, or because it requires data that is not available—complete observations of the contamination status of all nodes in the network.

A. L. Horn (✉)
Division of Epidemiology, Zoonoses and Antibiotic Resistance, Federal Institute for Risk Assessment (BfR), Berlin, Germany
e-mail: abbylhorn@alum.mit.edu

H. Friedrich
Kühne Logistics University, Hamburg, Germany

F. Ghanbarnejad et al. (eds.), *Dynamics On and Of Complex Networks III*,
Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-030-14683-2_7

Keywords Complex systems · Network source identification · Epidemic · Disease spreading · Network diffusion · Food supply networks · Foodborne disease

1 Introduction

An important problem for many networked systems involving spreading processes is identifying the source of the spreading agent; if the contaminated food source, patient zero, or the rumor originator is identified efficiently, damage can be prevented or reduced [7, 10, 13, 25].

Over the past couple decades there has been significant effort devoted to studying the dynamics of outbreaks on networks [5, 17, 19, 23, 24, 26, 31]; for a comprehensive review of epidemic spreading on complex networks, see [27]; for a review of information diffusion on complex networks including a comparative evaluation of available models and algorithms, see [36]. Most of this work has focused on the forward problem of understanding and forecasting the diffusion process and its dependence on the structure of the underlying network. However in recent years much work has emerged on the inverse problem of identifying the source of an outbreak spread in a network. This work covers problems in different contexts, including contagious disease infecting a human population; rumors or information diffusing through a social network; adoption of an idea, behavior change, or product in an organizational network; the spread of viruses on the internet; and the transport-mediated diffusion of contaminated individuals between cities. These contexts represent different spreading scenarios that require different modeling approaches for forward dynamics and inverse solutions.

Most studies of spreading processes in networks have been done in the context of epidemiology, modeling the spread of diseases or viruses through a host population. Network disease propagation models are based on the stages of disease as it infects individuals and spreads across contact links in a host population. Initially the entire population is susceptible to the disease; once any individual is exposed to an infectious contact they become infected and can infect others; from this point they can recover, be removed, become immune, or other variants. These models are referred to as compartmental models due to the disease compartments that individuals move between in illness progression: S—susceptible, I—infected, R—recovered or removed, etc.

Compartmental disease spreading models represent a simple contagion process, because only one direct contact with an infected neighbor is required for the contagion to be transmitted. Along with disease, information spread through a network has been shown to follow a simple contagion process. On the other hand, behavior change has been shown to spread as a complex contagion that requires multiple sources of exposure or reinforcement for the new behavior to be adopted.

A typical quantity that is studied in relation to network epidemic models is the epidemic threshold, or the set of conditions under which the disease will either proliferate or die out in the network. Unlike classical diseases or viruses spread through social contact networks, computer viruses have been shown to have an epidemic threshold of 0, meaning that the infectivity rate can be vanishingly small for the epidemic to happen. This is due to the scale-free structure of computer networks, which are extremely heterogeneous with a few nodes having an extremely high number of connections. The spread of computer viruses therefore diverges from classical diseases not due to the contagion model—both are simple contagion—but due to the heterogeneity of the network substrate over which computer viruses spread.

Another type of epidemic model is the metapopulation reaction–diffusion process, which in addition to contagion dynamics accounts for the role of movement or transport in diffusing a contamination in space. In this type of model, nodes represent subpopulations, such as cities, and links represent the movement of individuals between subpopulations. Individuals interact in each subpopulation according to assumptions of equal mixing or a local social network structure and disease spreads between these individuals according to a contagion model; this is the reaction process. The movement of individuals between subpopulations is the spatial diffusion process, often modeled over a network as a Markov transition process. Metapopulation models therefore depend both on the local social network structure at each node and on the spatial structure of the environment, transport infrastructures, traffic networks, and other movement patterns over which individuals diffuse.

Approaches to the source detection problem are developed in the context of one of these forward spreading processes. Most approaches have been devised in the context of simple contagion processes including infectious disease outbreaks in human contact networks or rumors spreading in social networks [1–3, 20, 33, 37, 38]. Another stream of work has focused on identifying the source of processes in which network-mediated spatial diffusion is the main vector of spread. This includes contagious diseases spread through drift in water systems [28] or spreading between cities by global air travel [6], and foodborne disease contamination spread through food distribution networks [14].

This article focuses on foodborne disease. The features that distinguish foodborne disease in the context of source identification have not previously been studied or identified. In this work we identify these features and conclude that most of the existing approaches to source detection cannot be implemented in the foodborne disease problem because they make assumptions about the transmission process that are unrealistic in the context of food supply networks—that is, identifying the source of an epidemic contagion [1–3, 20, 33, 37, 38] whereas foodborne contamination spreads through a transport network-mediated network diffusion process, or because it requires data that is not available—complete observations of the contamination status of all nodes in the network [8, 11, 30, 34] or timed network data [1–3, 15, 20, 21, 28, 33, 35]. We begin by first providing relevant background on outbreaks of foodborne disease and the contamination diffusion process.

1.1 Large-Scale Outbreaks of Foodborne Disease

The complexity and globalization of food production have made foodborne disease a widespread public health problem worldwide. A small but worrisome minority of outbreaks are generated by a contamination originating at the site of production or processing, generating a widespread diffusion of contamination through the supply chain and affecting a potentially great number of people across geographically distributed locations. As recent trends continue, including large-scale production practices and distribution over ever-larger distances, both the frequency and the severity of consequences of large-scale outbreaks are increasing. In the USA, the number of large-scale (i.e., multi-state) outbreaks increased by 135% in the years 1995–2004 to the years 2005–2014. These large-scale outbreaks accounted for 3% of total outbreaks, which includes localized, non-distributed incidents, but were responsible for 34% of hospitalizations and 56% of deaths [9].

During a large-scale outbreak of foodborne disease, rapidly identifying the source, including both the food vector carrying the contamination and the location source in the supply chain, is essential to minimizing impact on public health and industry. However, tracing an outbreak to its origin is a challenging problem due to the complexity of the food supply system. Furthermore, current investigation methods represent a missed opportunity to utilize valuable information to solve the source localization problem.

Food distribution is a complex system that can be seen as a network of trade flows connecting supply network actors. Identifying the source of an outbreak of contamination distributed across a network can best be solved by considering this network structure and the dimensions of information it contains. Together with reports of illness, this network information can be used to solve the problem of identifying the source of large-scale outbreaks.

To formulate the problem of source detection on a network, assumptions must be made regarding (1) the network and observation data available for source identification, and (2) the transmission process that led to the observations. Based on basic practical knowledge of food supply networks and the foodborne disease contamination process, in this article we introduce the source identification problem in the context of foodborne disease outbreaks and outline six features that distinguish this problem from source detection in other network contexts due to either practical data limitations or differences in transmission process mechanics. We then use the six features to categorize the existing literature on the network-based source detection problem according to relevance to the foodborne disease context.

2 Background and Definitions

2.1 Network-Based Source Identification

To solve the source detection problem in the context of foodborne disease, a network model of the supply of a specific food commodity is assumed as a given input. A probabilistic model of the transmission process of contamination spreading through this network is then postulated. In the following, we assume that a foodborne disease outbreak will originate from a single contamination source. This source sends out contaminated products that travel through the network according to the transmission model, resulting in observations of illness at a set of network nodes. The source identification objective is to minimize the error between the model-derived estimate of the location of the source and the true location of the source in the network, given the nodes associated with the observations of illness.

2.2 Food Supply Networks and Foodborne Disease Transmission

Food supply systems can be represented by a directed network structure consisting of multiple stages of production, distribution, storage, and consumption. Flows through the network are generally structured such that product is distributed in a forward direction along a *path*, or a collection of directed *edges* connecting supply *nodes* from origination to point of sale. A large-scale outbreak occurs when contaminated food departs from some source in an early stage of the network that is able to reach downstream nodes in geographically distributed locations. The contamination will eventually make its way to consumers, who develop illness some time after consuming the contaminated food. Case reports of illness are associated with the supply network node at which the offending product was purchased and *exits* the supply network, e.g. a retailer or restaurant; these nodes can be considered *infected*.

The network in Fig. 1 represents a supply network in which contamination at a food producer has spread through the supply network, leading to reports of illness at three different retailers. With this structure mapped, it is straightforward to utilize all case data (i.e., evidence) available during an event to identify the set of *feasible* sources of contamination, that is, the set of nodes that connect to all known contaminated nodes. Network structural information thus provides a first cut into the source identification problem by enabling the identification of feasible sources. To differentiate between the feasible sources, further dimensions of information available within the network can be leveraged. Each edge contains information about the volume of goods traded between supply network actors. Volume-weighted information is a source of heterogeneity that can be thought of as the relative propagation potential of a given edge, providing insight into the paths along which contaminated product is likely to have traveled.

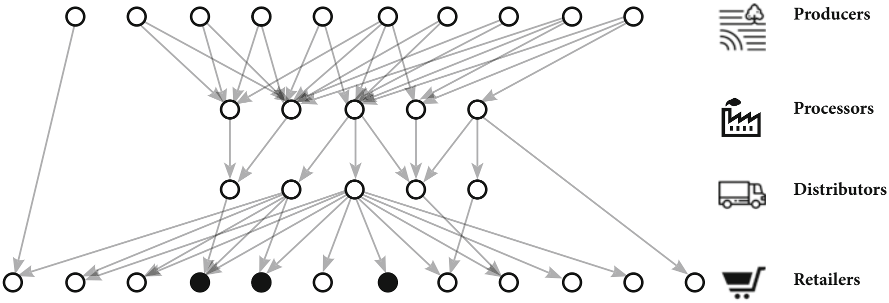

Fig. 1 Illustration of a food distribution network with three reported cases of illness (at the shaded nodes) linked to retailer nodes. Figure source: [8]

3 Distinguishing Features of Foodborne Disease Transmission

3.1 A Transport, Not Epidemiological, Transmission Process

Many network-based source detection methods are designed to identify the source of an infectious contagion. These methods often assume some variant of the epidemiological model of contagion transmission, including the widely used susceptible-infected (SI) or susceptible-infected-recovered (SIR) models. However the transmission of contamination through the food supply to people is different from the disease contagion process from people to people. Contamination spreads as contaminated (solid, perishable) food moves through the supply network after being inoculated by the pathogen at the source. As the food is transported through the supply network, the pathogenic quantity will generally remain conserved, meaning it will neither spread to other food items nor decay significantly in infectivity [18, 29]. The former is due to a number of factors including the lack of contact between packaged items, the lack of interaction or mixing between unpackaged items, and the biological insusceptibility of contamination to transmission and decay, i.e. low infectivity and recovery rates.

Due to this conservation of contamination, the spreading process in the context of foodborne disease primarily involves the contaminated food being spatially distributed along the network without decaying (i.e., recovery) or growing (i.e., infection) the contamination along the way. Contaminated food items cause infection in people when the food is consumed, but this process does not represent a classical infection dynamics because the contamination is directional (food to human) and largely does not spread between people. Contagion processes represent a different dynamics; if these are applied to the foodborne disease situation, the extremely low infection rate would mean that when individual food items come into contact, the infection will not be transmitted and will die out. The diffusion along the network is the mechanism that moves the contamination forward through the supply chain. To reflect these diffusion dynamics, the foodborne disease contamination spreading process has therefore been modeled as a simple Markov transmission process [14].

It would be possible to model the foodborne disease spreading process using a metapopulation reaction–diffusion process, as discussed in Sect. 1, where nodes represent locations in the supply chain containing a constant "subpopulation" of food items, and links represent the transport of food items between supply chain locations. However because contagious transmission is largely not occurring between food items, a metapopulation model would add more complexity (by incorporating the inactivated local contagion process along with the diffusion process) without incorporating more of the dynamics of the spread of contamination by food through the supply chain. Therefore in the following, we will refer to the foodborne disease contamination process as a diffusion-type process by which we mean exclusively network-mediated diffusion and not contagion.

Finally, the observation data available for source identification occurs on the human level and not on the food item level, and only via infection status, (I) in the SI/R model. Observations of contamination occur when people report illness. Each illness is linked to a supply network node at which the contaminated food was purchased. Data regarding the contamination status of individual food items is not normally available during an investigation. Furthermore, it is not possible to establish from the illness reports whether a supply node has ever received contaminated food and is thus susceptible (S), as it may have led to illnesses that went unreported. Methods that rely on observations of susceptible status or that assume nodes not reporting infection are contamination-free (also called "negative information") are thus non-applicable in this setting.

3.2 Observations are Sparse

Though the contamination will travel through multiple network nodes on its journey through the supply network, it is only observed when illness is reported in connection with the exiting or *absorbing* node at which contaminated food was purchased. The contamination status of *transient* nodes involved in the production, processing, or storage of food, though closer to the source in number of network edges, will remain hidden to investigators unless further investigations are performed (normally during later stages of an investigation). Furthermore, even at the consumption level, the overwhelming majority of foodborne illness cases are either not identified or logged by authorities, with official estimates of underreporting varying from 10 to 75 times for different pathogens [32]. A trivial implication of the sparsity of observations is that it is unrealistic to assume, as some source detection methods do, that the contamination status of all nodes in the network is known.

3.3 Observations will Always be Spaced Far from the Source

The placement of observations only at absorbing nodes also means that there will be a large network distance between the source and each observation, increasing the number of possible paths that could have been traveled and in turn the uncertainty in the *structure* of the diffusion trajectory. At the same time, the differing volume-weights along the edges of the supply network provide valuable information for inference. Given the large uncertainty in the diffusion structure, approaches to source detection that consider network structure alone will be inferior to those that consider this weighted information.

3.4 Similar Path Lengths

Due to the staged structure of the food supply network, paths through the network from source to observation will be close to the same length in terms of number of network edges. This is common for supply chain networks of all types, and can be observed in models of food supply networks across all product groups [4, 12]. Many existing source detection methods simplify the inference process by assuming that the contamination traveled across the shortest path from the source to each observation, or otherwise by leveraging shortest path properties of graphs. These approximations will apply poorly in the food supply network context where most paths will be indistinguishable in length.

3.5 Multiple Candidate Paths

Between any possible source and observation in a food supply network, there exist multiple paths of travel of similar weight or likelihood. This is due to the lack of monopolies in food production, trade, and retailing markets: any given food type will be distributed through multiple larger retailers or wholesalers, each dealing with similarly large volumes of product [4, 12]. Certain source detection methods make the simplifying assumption that the contamination travels across the single highest-probability path between a source and observation. These methods will be inaccurate in the food supply network setting where transmission dynamics are not necessarily dominated by a small percentage of connections.

3.6 Data on Times Through the Network are Lacking

In theory, there should be a signal for source detection from the timed reports of illnesses combined with a model of the time it takes to transmit the contamination. Each collection of edges in a network path encodes information about the time delay that a contaminated product could have taken to travel these steps. These delays will be distributed differently according to parameters like the distance and speed of travel and supply network logistics encountered. However, there is significant temporal uncertainty in the contamination transmission process. The time the contamination may spend in storage, both at various nodes along the supply network (e.g., warehouses) and with the consumer after purchase, as well as during the incubation period, can be significant and vary widely—and potentially much more so than the time spent in travel. Furthermore, while the times of infection are available to some degree of accuracy (recorded according to patient recalled time of illness onset), data on storage times through the network are unavailable with the exception of a few case-specific customer or retailer survey studies [18, 29]. Therefore, while time can be an important aspect in some foodborne disease source detection applications, time-based methods are not currently implementable in the foodborne disease context given available data.

4 Categorization of Literature

Many approaches to the network source detection problem have been developed in recent years, though none of these methods have specifically considered the context of outbreaks of foodborne disease. We now review the major themes in the existing work, using the features described above to guide the discussion in terms of relevance to the problem on food supply networks. The categorization of existing work in terms of these features is summarized in Table 1.

The earliest approaches to source detection are based on complete observations, relying on knowing the contamination status (SI/R) of each node in the network at a fixed point in time [8, 11, 30, 34]. These methods do not incorporate information about differing weights along edges but are based solely on graph structure by employing notions of network centrality, the intuition being that the node most "central" to the observed contamination process is the source. The seminal work by Shah and Zaman [34] introduces the measure of *rumor centrality*, which considers the number of linear extensions between each source and the infected nodes. The method and analytical results concerning detection probability are derived for trees or tree-like graphs; to apply to general networks, a Breadth-First-Search (BFS) heuristic that assumes the contamination traveled across the shortest paths to the observations must be used. Other methods based on *betweenness centrality* [8] and *eigenvector centrality* [11, 30] apply to general networks without employing a shortest path heuristic, although the calculation of betweenness is based on shortest path properties. These methods were important for establishing foundational results on the network source detection problem but are impractical for real network-outbreak scenarios due to the complete observation assumption.

Many methods have since been developed for the more realistic setting that only a subset of the contaminated nodes are observable, i.e. partial observations. These can be categorized into temporal methods—approaches designed to make use of

Table 1 Categorization of existing work on the source detection problem according to relevance to the foodborne disease context

Source identification methodology of existing work	Limitations of source identification methodologies in foodborne disease context					
	(1)	(2)	(3)	(4)	(5)	(6)
	Only SI/R (no diffusion component)	Assumes complete observations	Ignores weights	Only shortest paths	Only dominant paths	Assumes times through network
Rumor centrality [34]		X	X	X		
Betweenness centrality [8]		X	X	X		
Eigenvector centrality [11, 30]		X	X			
Message passing [20]	X					X
Belief propagation [2]	X					X
Analytic combinatoric [3]	X					X
Gaussian [21, 28]				X		X
Four-metric [33]	X			X		X
Monte Carlo [1]	X			X		X
Analytical time-varying networks [15]				X		X
Time-reversal backward spreading [35]						X
Jordan centrality [37, 38]	X			X		
Effective distance [6, 22]					X	
Multiple-paths diffusion [14]						

the information from the timed reports of illness and times through the network, and non-temporal methods—approaches that rely only on the node location where contamination has been reported. The temporal category includes methods assuming discrete-time epidemic (SI/R) contagion models based on *dynamic message-passing* [20], *Bayesian belief propagation*, [2], *analytic combinatoric* approaches [3]. The *analytic-combinatoric* method [3] builds on the approach of [20] and [2] by removing the node-independence assumption of [20] and the tree-like contact network assumption that both [20] and [2] are predicated on to compute the exact source probability distribution for general contact network structures. Because the analytical calculations increase exponentially for non-tree-like networks, a computationally feasible Monte Carlo estimation approach is provided and demonstrated empirically to provide comparable results with the analytic method. The approach of [3] applies both to static and temporally evolving networks.

A separate approach involves continuous-time *Gaussian* transmission models [21, 28]. While a continuous-time transmission model is a better approximation for realistic settings, the approach in [21, 28] is limited by being designed for trees and extended to general graphs via a BFS (shortest-path) heuristic. Other temporal methods have been proposed that observe the contamination status of a subset of sensor nodes at user-controlled intervals invoking a Four-Metric approach [33], Monte Carlo methods [1], or analytical methods for time-varying networks [15]. A separate approach is based on time-reversal backward spreading, where link weights are set equal to travel time and not spreading propensity [35]. These methods are impractical for the foodborne disease context given the lack of temporal data on times through the network available for solving the problem, as discussed in Sect. 3.6.

Fewer approaches to source detection exist within the category of non-temporal approaches based on partial observations. A line of work based on the notion of *Jordan centrality* has led to multiple variants of a technique that chooses the source node with the shortest maximum path length over all observations, that is, the Jordan center [37]. While this method has been extended to incorporate weights along the edges[1] [38], it relies on path lengths to discriminate between sources. Furthermore, the technique is designed for tree-like networks; for application to general topologies an alternate procedure based on closeness centrality (i.e., counting the sum of the shortest path to each observation) is proposed.

In addition, many of the methods based on partial observations in both the temporal and non-temporal categories are developed in the context of contagion spreading models [1–3, 20, 33, 37, 38], and are therefore inapplicable in the case of the supply network-mediated diffusion process of foodborne disease spread. As explained in Sect. 3.2, network-based diffusion is the mechanism moving contamination forward through the supply chain, which represents a different dynamics than contaminated individuals changing infection state and growing the infection. If contamination

[1] In the contagious disease context, normalized weights can be interpreted as heterogeneous infection probabilities.

models are applied to foodborne disease spread, the extremely low infectivity rate and recovery rates would mean that the disease would die out, and the forward diffusion of already-contaminated items would not be accounted for.

Another line of work in the category of non-temporal approaches involves a measure of *Effective Distance* on a network [6]. The Effective Distance method is developed for identifying the source of infectious disease outbreaks spreading through global mobility networks and is therefore devised in the framework of metapopulation reaction–diffusion models. However, it does not depend explicitly on the infection quantities, but only on flow transitions between nodes. It is therefore applicable to network-diffusion-only type processes such as foodborne disease and has been evaluated in application to the 2011 outbreak of EHEC in sprouts [22].

The method is based on the concept that the trajectory of a particle diffusing through a network will primarily follow the shortest, highest probability path to any other node. The true source of an outbreak should therefore be the node that exhibits the set of shortest, highest probability paths to the outbreak node set. Based on this logic, the authors introduce a metric for the Effective Distance $d_{eff}(i, j)$ between two connected nodes i and j, defined such that the likelier the connection, the shorter the Effective Distance. This is given as

$$d_{eff}(i, j) = 1 - \log p_{ij}, \tag{1}$$

where p_{ij} is the probability of transiting from i to j. The effective length of a given path γ_{so} between source node s and observation node o is then defined to be the sum total of the Effective Distances of each edge $(i, j) \in \gamma_{so}$. As discussed, the concept of [6, 22] is to focus on the shortest Effective Distance path over all possible paths $\gamma_{so} \in \Gamma_{so}$ from s to o. The Effective Distance between s and o is then defined as

$$\begin{aligned} D_{eff}(s, o) &= \min_{\gamma_{so} \in \Gamma_{so}} \sum_{(i,j)\in\gamma_{so}} 1 - \log p_{ij} \\ &= \min_{\gamma_{so} \in \Gamma_{so}} [|\gamma_{so}| - \log P(\gamma_{so}|s)]. \end{aligned} \tag{2}$$

The Effective Distance of a path therefore results from a multifactorial objective function that penalizes topologically long path lengths (the $|\gamma_{so}|$ term in the minimization) while rewarding high path probabilities (the $-\log P(\gamma_{so}|s)$ term). To identify the source of an outbreak, the single shortest Effective Distance path to each observation is identified. The source is then chosen as the node that minimizes the average and variance of the shortest Effective Distance path to each observation.

As mentioned above, the Effective Distance method was designed for application to infectious disease outbreaks spreading over global mobility networks. These networks are characterized by great heterogeneity in path lengths and probabilities, meaning that spreading processes on these networks will be dominated by a small percentage of the shortest, highest probability transport connections. As expected, the Effective Distance method performs well in settings involving outbreaks of infectious disease (e.g., SARS, H1N1) spreading through global air travel networks

[6]. Nonetheless, it is a heuristic approach that considers only a single path to each observation. While this type of approximation may be justified in certain network contexts such as the global air mobility networks the method was designed for, it is not adapted for the structure of food supply networks which are characterized by homogeneity in path lengths and the existence of multiple paths of similar probability (see Sect. 3). When the method is applied to the 2011 EHEC (foodborne disease) outbreak, source identification results are less accurate and more unstable than the infectious disease case examples [6, 22].

Two recent works have addressed the single path limitation. First, Ianelli et al. [16] have developed a generalization of the Effective Distance approach to include multiple transmission routes in estimating disease arrival times. This work leverages random walk theory to analytically demonstrate that the single path approach is an approximation of more general logarithmic network-based measures. While both methods are developed in the framework of metapopulation reaction–diffusion models, only the multiple paths approach depends explicitly on the dynamical quantities of the SIR model. This generalized effective distance approach for estimating (forward) disease propagation arrival times is therefore a departure point for an improved and analytical approach to the (inverse) source detection problem for metapopulation propagation processes like infectious diseases spreading through global air traffic networks.

More recently, the source detection problem for network-diffusion-only type processes such as foodborne disease has been solved using a similar analytical approach to account for all trajectories between source and observation. The work of Horn and Friedrich [14] formulates a probabilistic model of the contamination diffusion process as a random walk on a network and derives the maximum likelihood estimator for the source location. By modeling the transmission process as a random walk, this work develops a novel, computationally tractable solution to the inverse problem that accounts for all possible paths of travel through the network. Improvements in accuracy and stability are demonstrated in comparison with the single paths approach of [6, 22], when both methods are applied to different network topologies including stylized models of food supply network structure as well as the 2011 EHEC outbreak in Germany.

5 Summary

Many existing approaches to the source detection problem cannot be implemented in the foodborne disease context because they are designed for a different purpose—identifying the source of an epidemic contagion [1–3, 20, 33, 37, 38] whereas foodborne disease is spread according to a network-mediated diffusion process, or because they require data that is not realistically available—complete observations of the contamination status of all nodes in the network [8, 11, 30, 34] or timed network data [1–3, 15, 20, 21, 28, 33, 35]. Those that are implementable are limited by unrealistic assumptions regarding the transmission process. These methods apply tree-like approximations to deal with general graphs, assuming contamination always travels from source to observations along the shortest, highest probability paths [6, 22]. While this type of approximation is justified in certain network contexts, food supply networks are not well approximated by tree structure. Moreover, these methods are by definition approximations that do not explore the full set of trajectories between each source and observation.

To address this limitation, recent work has developed a source detection approach based on a random walk transmission model that presents a computationally tractable approach to calculate the total probability of traveling between a source and each observation along all possible paths of all possible lengths [14]. The resulting approach is not only relevant for solving the source identification problem in food supply networks but also represents a methodological improvement for source identification in diffusion processes more generally.

References

1. Agaskar, A., Lu, Y.M.: A fast Monte Carlo algorithm for source localization on graphs. In: Proc. SPIE Optical Engineering Applications, San Diego, CA (2013). Art. no. 88581N
2. Altarelli, F., Braunstein, A., Dall'Asta, L., Lage-Castellanos, A., Zecchina, R.: Bayesian inference of epidemics on networks via belief propagation. Phys. Rev. Lett. **112**(11), 118701 (2014)
3. Antulov-Fantulin, N., Lančić, A., Šmuc, T., Štefančić, H., Šikić, M.: Identification of patient zero in static and temporal networks: robustness and limitations. Phys. Rev. Lett. **114**(24), 248701 (2015)
4. Balster, A., Friedrich, H.: Dynamic freight flow modelling for risk evaluation in food supply. Transport. Res. E Log. **121**, 4–22 (2019)
5. Brockmann, D., David, V., Gallardo, A.M.: Human mobility and spatial disease dynamics. Rev. Nonlinear Dyn. Complex. **2**, 1–24 (2010)
6. Brockmann, D., Helbing, D. The hidden geometry of complex, network-driven contagion phenomena. Science **342**(6164), 1337–1342 (2013)
7. Colizza, V., Barrat, A., Barthelemy, M., Valleron, A.J., Vespignani, A.: Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med. **4**, 1–16 (2007)
8. Comin, C.H., da Fontoura Costa, L.: Identifying the starting point of a spreading process in complex networks. Phys. Rev. E **84**(5), 056105 (2011)
9. Crowe, S.J., Mahon, B.E., Vieira, A.R., Gould, L.H.: Vital signs: multistate foodborne outbreaks – United States, 2010–2014. MMWR Morb. Mortal. Wkly. Rep. **64**, 1221–1225 (2015)
10. Finkelstein, S.N., Larson, R.C., Nigmatulina, K., Teytelman, A.: Engineering effective responses to influenza outbreaks. Ser. Sci. **7**, 119–131 (2015)
11. Fioriti, V., Chinnici, M.: Predicting the sources of an outbreak with a spectral technique. arXiv preprint arXiv:1211.2333 (2012)
12. Friedrich, H.: Simulation of Logistics in Food Retailing for Freight Transportation Analysis. Doctoral dissertation. Karlsruhe Institute for Technology, Karlsruhe (2010)
13. Hollingsworth, T.D., Ferguson, N.M., Anderson, R.M.: Will travel restrictions control the international spread of pandemic influenza? Nat. Med. **12**, 497–499 (2006)
14. Horn, A.L., Friedrich, H.: Locating the source of large-scale outbreaks of foodborne disease. J. R. Soc. Interface **16**, 20180624 (2019). https://doi.org/10.1098/rsif.2018.0624

15. Hu, Z.L., Shen, Z., Cao, S., Podobnik, B., Yang, H., Wang, W.X., Lai, Y.C.: Locating multiple diffusion sources in time varying networks from sparse observations. Sci. Rep. **8**(1), 2685 (2018)
16. Iannelli, F., Koher, A., Brockmann, D., Hšvel, P., Sokolov, I.M.: Effective distances for epidemics spreading on complex networks. Phy. Rev. E **95**(1), 012313 (2017)
17. Keeling, M.J., Eames, K.T.D.: Networks and epidemic models. J. R. Soc. Interface **2**, 295–307 (2005)
18. LeBlanc, D.I., Villeneuve, S., Beni, L.H., Otten, A., Fazil, A., McKellar, R., Delaquis, P.: A national produce supply chain database for food safety risk analysis. J. Food Eng. **147**, 24–38 (2015)
19. Lind, P.G., da Silva, L.R., Andrade, J.S., Herrmann, H.J.: Spreading gossip in social networks. Phys. Rev. E **76**, 036117 (2007)
20. Lokhov, A.Y., Meézard, M., Ohta, H., Zdeborovaá, L.: Inferring the origin of an epidemic with a dynamic message-passing algorithm. Phys. Rev. E **90**(1), 012801 (2014)
21. Louni, A., Subbalakshmi, K.P.: A two-stage algorithm to estimate the source of information diffusion in social media networks. In: Proceeding IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, pp. 329–333. IEEE, Piscataway (2014)
22. Manitz, J., Kneib, T., Schlather, M., Helbing, D., Brockmann, D.: Origin detection during foodborne disease outbreaks-a case study of the 2011 EHEC/HUS outbreak in Germany. PLoS Curr. **6** (2014)
23. Moore, C., Newman, M.E.J.: Epidemics and percolation in small-world networks. Phys. Rev. E **61**, 5678–5682 (2000)
24. Newman, M.E.J.: Spread of epidemic disease on networks. Phys. Rev. E **66**, 016128 (2002)
25. Nigmatulina, K.R., Larson, R.C.: Living with influenza: impacts of government imposed and voluntarily selected interventions. Eur. J. Oper. Res. **195**(2), 613–627 (2009)
26. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Phys. Rev. Lett. **86**, 3200–3203 (2001)
27. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., Vespignani, A.: Epidemic processes in complex networks. Rev. Mod. Phys. **87**(3), 925 (2015)
28. Pinto, P.C., Thiran, P., Vetterli, M.: Locating the source of diffusion in large-scale networks. Phys. Rev. Lett. **109**(6), 068702 (2012)
29. Pouillot, R., Lubran, M.B., Cates, S.C., Dennis, S.: Estimating parametric distributions of storage time and temperature of ready-to-eat foods for US households. J. Food Prot. **73**(2), 312–321 (2010)
30. Prakash, B.A., Vreeken, J., Faloutsos, C.: Efficiently spotting the starting points of an epidemic in a large graph. Knowl. Inf. Syst. **38**(1), 35–59 (2014)
31. Riley, S.: Large-scale spatial-transmission models of infectious disease. Science **316**, 1298–1301 (2007)
32. Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R.V., Widdowson, M.A., Roy, S.L., Jones, J.L., Griffin, P.M.: Foodborne illness acquired in the United States – major pathogens. Emerg. Infect. Dis. **17**(1), 7–15 (2011)
33. Seo, E., Mohapatra, P., Abdelzaher, T.: Identifying rumors and their sources in social networks. In: Proceeding SPIE Defense, Security, and Sensing, Baltimore, MD, USA (2012)
34. Shah, D., Zaman, T.: Rumors in a network: who's the culprit? IEEE Trans. Inf. Theory **57**(8), 5163–5181 (2011)
35. Shen, Z., Cao, S., Wang, W.X., Di, Z., Stanley, H.E.: Locating the source of diffusion in complex networks by time-reversal backward spreading. Phys. Rev. E **93**(3), 032301 (2016)
36. Zhang, Z.K., Liu, C., Zhan, X.X., Lu, X., Zhang, C.X., Zhang, Y.C.: Dynamics of information diffusion and its applications on complex networks. Phys. Rep. **651**, 1–34 (2016)
37. Zhu, K., Ying, L.: Information source detection in the SIR model: a sample path based approach. In: Proceedings of Information Theory Applications Workshop (ITA), San Diego, CA, pp. 1–9. IEEE, Piscataway (2013)
38. Zhu, K., Ying, L.: A robust information source estimator with sparse observations. Comput. Soc. Net. **1**(1), 1 (2014)