LOCATING THE SOURCE OF
LARGE SCALE OUTBREAKS OF FOODBORNE DISEASE

by

Abigail Lauren Horn

B.A. Physics
College of Creative Studies, University of California, Santa Barbara, 2007

Submitted to the Institute for Data, Systems, and Society
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Engineering Systems

at the

Massachusetts Institute of Technology

September 2016

Signature redacted

Author...
Institute for Data, Systems, and Society
August 22, 2016

Certified by....................... Signature redacted ....
Richard C. Larson
Mitsui Professor, Data, Systems, and Society
Thesis Supervisor

Certified by.......... Signature redacted ....
Stan N. Finkelstein
Senior Research Scientist, Engineering Systems
Thesis Co-Supervisor

Certified by...... Signature redacted ....
Marta C. Gonzalez
Associate Professor, Civil and Environmental Engineering
Committee Member

Certified by.................... Signature redacted ....
Hanno Friedrich
Assistant Professor of Freight Transportation - Modelling and Policy, Kühne Logistics University
Committee Member

Accepted by................. Signature redacted ....
John Tsitsiklis
Clarence J. Lebel Professor, Electrical Engineering; Graduate Officer, Institute for Data, Systems, and Society

LOCATING THE SOURCE OF
LARGE SCALE OUTBREAKS OF FOODBORNE DISEASE

by

Abigail Lauren Horn

Submitted to the Engineering Systems Division
On September 2, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Foodborne illness outbreaks impose enormous health and economic burdens in the US. Identifying the origin of the contaminated food causing an outbreak is a challenging problem due to the complexity of the food supply and the absence of coherent labeling and distribution records. Current investigative methods are slow, resource intensive, and the overwhelming majority of investigations are unsuccessful in identifying the location source of an outbreak. New tools and approaches that take advantage of modern data and analytical techniques are needed to more quickly identify outbreak origins and prioritize response efforts.

The practical objective of this work is to improve the food safety regulator's ability to efficiently locate the source of an outbreak while contamination-caused illnesses are occurring, thereby resolving investigations earlier and averting potential illnesses. This thesis develops new methods that leverage currently unutilized or underutilized sources of information to identify the location source of an outbreak. A novel, network-theoretic approach to source detection is developed that (1) immediately identifies all *feasible* source locations, (2) ranks the *feasible* locations by the likelihood that each one is the *true* source, and (3) develops a decision model for guiding investigators to implement effective interventions. The approach functions on food system network data, reported cases of illness at specific times and locations, and a prior probability function over likely sources. The methodology is the first to be designed specifically for tracing back outbreaks on food distribution networks.

A Monte Carlo simulation environment was developed to evaluate traceback performance and robustness across a wide range of network structures and outbreak scenarios. When compared against existing traceback methods, both those currently in practice and those in academic literature, this methodology demonstrates significant improvements in accuracy, efficiency, and speed. Specific results suggest the approach can provide substantial benefits to the investigation process by identifying the source early enough in an outbreak's progression that a substantial fraction of cases of illness can be averted. These computational results serve as a powerful first step towards validating the accuracy and applicability of the approach. The immediate next step will be to demonstrate accuracy when applied to real food distribution networks. While acquiring representative network data for this purpose presents significant practical challenges, an additional contribution of this work is the identification of a representative network model that can be integrated with the source identification methodology, forming a *holistic traceback framework*.

Thesis Supervisor: Richard C. Larson
Title: Mitsui Professor, Data, Systems, and Society

Acknowledgements

I first and foremost would like to share my profound gratitude to my advising team, Richard Larson and Stan Finkelstein. Dick and Stan have been incredibly generous in co-advising and sheparding me through my PhD. Dick has been an incredible mentor and his guidance has helped to shape not only the course of my academic career, but also the way I see the world. A large part of the PhD process is deciding what problem to focus on. Dick has guided me, by example, on developing a type of problem that is both interesting and has a high potential for impact. His teaching and guidance are led by a desire to make challenging concepts accessible, and more broadly, to bring knowledge and learning to all through his activities both at MIT and beyond, and I will seek to carry these principles forward. Stan has been a terrific advisor. Since we met in Portugal in 2010, he has played a pivotal role in my path, guiding me through a deeper reach into my PhD while always reminding me to consider and be led by the broader applications of this work. I will be forever grateful to both Dick and Stan for mentoring me through these significant milestones, and for providing so many important opportunities along the way.

Hanno Friedrich has been a wonderful advisor and has played a critical role on my committee, demonstrating the relevance of this dissertation in other fields of research, developing extensions of this work, and guiding me to make the transition to next steps in its validation. I am very grateful to have had the opportunity to learn from Marta Gonzales in many ways: in class, in conversation, and by example – from the exciting research field she is helping to create. I greatly appreciate Marta's support and guidance.

This work is the product of the ESD doctoral program, and I'm incredibly grateful to have benefited from the ideals and curriculum of this unique department. In particular, the engagement and interest of professors such as Joe Sussman, Chris Magee, Oli de Weck, and Jessika Trancik have helped to focus and frame this research, and has served to reaffirm the value of my work, for which I am deeply appreciative.

I would like to thank the food safety regulators and investigators who took the time to discuss and advise this work. In particular, a series of meetings with Sherri McGarry, Senior Advisor with the Coordinated Outbreak Response and Evaluation group at the FDA, and Erica Pomeroy, Consumer Safety Officer at the FDA San Francisco District Office, to discuss the real world difficulties of foodborne disease outbreak traceback investigations, were particularly influential in the framing of this dissertation.

This work would not have been possible without the belief, support, and funding of the Robert Wood Johnson Foundation (RWJF), administered by the National Coordinating Center for Public Health Services and Systems Research (PHSSR), as well as the assistance and efforts of their excellent team of dedicated staff.

This work has benefitted enormously from the work of two extremely bright MIT undergraduates, who I had the opportunity to work with as part of MIT's UROP program. Vinati Kaul helped to determine the direction of this work, contributing her ability to extract key information from broad sources. Elena Polozova has contributed in a very material way to making this work possible. Her abilities, ingenuity, and curiosity to learn have led her to come up with interesting solutions to complex challenges.

I need to thank the incredible peer group of ESD colleagues and friends, from whom this work has benefitted immensely. I would like to extend my gratitude in particular to Alex Jacquillat, John Helferich, Vivek Sakrahni, and Maite Pena-Alcaraz, all of who have shared their time to advise me in structuring my thoughts and the presentation of my work. I am extremely grateful for the guiding role they have played in my PhD process, in addition to their friendship. I also need to thank the incredibly supportive staff at ESD, including Elizabeth Milnes who watches over us all as our "guardian advisor," and Frances Marrone, who has provided the most thoughtful assistance in so many critical moments.

Finally, I need to thank my incredible and very unique family. My beautiful, world-embracing parents, who have created the most exciting, interesting world for me they possibly could have, and for instilling the importance of following a path of equal parts determination, curiosity, and passion. My incredibly generous and supportive sisters, who have provided me with a limitless support of confidence. My late father in law, who set an example for living a life in pursuit of knowledge, and in recognizing the importance of the process. And last but most of all to my mewgnificent husband, who entirely made this whole process possible, filling every role imaginable to support me, enduring and overcoming the many challenges of this process to see the most in every activity and stage, and continuing to impress me by his equally important traits of boundless capabilities and positivity.

Contents

Executive Summary

The impact of large-scale, multi-state outbreaks of foodborne disease

The complexity and globalization of food production have made foodborne disease a widespread public health problem in both developed and developing countries. The Center for Disease Control (CDC) estimates that each year in the U.S. 48 million illnesses, 128,000 hospitalizations, and 3000 deaths result from foodborne disease, at an estimated annual loss of $152 billion (Osterholm 2011). Due to underreporting or unknown agents, these numbers are potentially much higher (Batz 2005, Batz and Morris 2012, Scallan et al. 2011). Most foodborne outbreaks involve a source of contamination at the point of preparation or sale and affect a small group of people in a localized area. However 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 much greater number of people across geographically distributed locations. As recent trends continue, including large-scale production practices and distribution over ever-larger distances, both the prevalence and the severity of consequences of these major outbreaks are increasing. From 2005 – 2014, nearly 200 multi-state outbreaks were identified and investigated in the US as compared with 85 over the years 1995 – 2004; these multi-state outbreaks accounted for 3% of total outbreaks, but were responsible for 34% of hospitalizations and 56% of deaths (Crowe et al. 2015). Since the probability of any given food generating a large, multi-state outbreak is extremely low but the impact is potential very great, these are classic low-probability, high-consequence events.

Current research efforts and federal initiatives to mitigate the impact of foodborne disease outbreaks, including the landmark 2012 Food Safety Modernization Act (FSMA), have focused largely on addressing the causes of outbreaks (Nuzzo 2013). While these efforts are essential to minimizing the risk of illness associated with the consumption of contaminated produce, they do not provide the tactical support necessary for response to foodborne illness outbreaks that occur when such preventive efforts have failed (IOM 2010). As a result, despite the dedicated efforts of food safety officials across the country, our current capacity to identify the origin of illness in multi-state outbreaks is characterized by typical timelines of 1-2 months. Investigations are completed in many cases after the outbreak has ended and the contamination has made its way through the supply chain, meaning that no cases of illness are averted as a result. Furthermore, the majority of outbreaks remain unsolved, meaning that the food and/or location sources of the outbreak are never identified and measures to reduce the impact of the outbreak could not be implemented (Wilkins et al. 2015, McEntire and Bhatt 2013). In the 13,352 foodborne disease outbreaks (causing 271,974 illnesses) documented by the CDC during 1998-2008, only 4,887 (37%) were traced back to a single food vehicle and pathogenic source, with less than 15% of these to a specific contamination point (Painter et al. 2013).

Tactical response to large-scale outbreaks

This thesis focuses on improving the response to large-scale, multi-state outbreaks. There are three standard components to the outbreak response and investigation process, beginning from the time the first case presents symptoms and ending when the contaminated product has been conclusively identified: (i) detecting that an outbreak is occurring, (ii) identifying the food vector causing the outbreak, and (iii) identifying the location source of the outbreak at a farm or processing center. There are multiple opportunities for improving each component of the outbreak investigation process that can have positive and meaningful impacts on public health. Novel strategies facilitated both by new tools (e.g., "next generation sequencing") and the revolutionary availability of digital sources of data related to food sales and consumption trends are being developed to contribute to the ability to (i) detect outbreaks and (ii) implicate the food vector causing the disease. However, methodologies that harness these new tools and data sources to contribute to solving part (iii) of the outbreak investigation, localizing the source, have not been brought to bear on the problem, despite the availability of a large amount of relevant system information, as we discuss below.

To fill this gap, this thesis develops novel methods that leverage currently unutilized or underutilized sources of information to contribute to the ability to identify the location source of an outbreak. If public health officials can more quickly and successfully identify the location source of contamination, the outbreak can be stopped from spreading, the number of people who get ill can be reduced, and unmerited damages to the food industry can be avoided. This thesis therefore takes a deep look into the question, *how can we locate the source of contamination with greater accuracy, certainty, and speed?*

Beyond its immediate and clear public health relevance, improving the ability to locate the outbreak source has real application to a current federal-level policy debate. Among the many provisions of the landmark 2011 Food Safety Modernization Act (FSMA) is an extension of traceability requirements. Specifically, FSMA extends the authority of the FDA to establish, as appropriate, "a product tracing system to *receive information that improves the capacity to effectively and rapidly track and trace food* that is in the United States or offered for import into the United States." The Act does not include specific provisions for implementing the law, and the regulatory timeline for determining the requirements is still developing. We suggest that the source localization framework developed in this thesis presents a viable approach to improving the capacity to "effectively and rapidly track and trace food."

Regulatory approach to source localization

The current regulatory approach generally involves "triangulation," or tracing back the unique distribution paths of products from several locations to determine if there is a common point of convergence in the supply chain, such as a common date and location of harvest or place of manufacture (FDA 2001, Wilkins et al. 2015). Investigators will choose cases that are part of distinct sub-clusters of contamination emerging at different restaurants or retailers, ideally starting from locations that are geographically distant from one another, e.g. a case in Los Angeles, Houston, and Boston. For each case, investigators will start from the retail establishment where the offending product is known to have been purchased or consumed, and trace the product back through each step of the supply chain, first determining the set of logistic service providers who could have brought the product to the retailer, then the set of processors who supplied to those logistic service providers, and so on, until the set of raw food production locations have been identified. They will then compare the supply chain actors uncovered as being potentially connected to each traced-back case, looking for commonalities. Because of the complexity of the supply chain and the large number of possible pathways to collect data for, the process is time and resource intensive. According to an estimate provided by investigators with the Minnesota Department of Public Health, 8 to 24 person hours are required to collect paperwork and create a product trace diagram for 1 to 2 contamination cases (Smith 2015).

The triangulation approach is fundamentally limited by its inability to leverage available information that can be contribute to identifying the location source of an outbreak. Because of the logistical limitations, investigators are only able to make use of a small subset of the reported cases of illness – data that serves as evidence in the source location problem. With only a few pieces of evidence, the time consuming traceback will often be unsuccessful in narrowing down the problem significantly. Furthermore, the process considers only supply chain structural information, that is, whether a link exists between different supply chain actors, while missing other dimensions of supply chain data that can help to differentiate between possible sources.

Moreover, triangulation represents a missed opportunity to utilize valuable system information to solve the source localization problem. When each supply chain pathway is considered individually, the greater food distribution system that these pathways and their related supply chain actors are a part of is ignored.

Network approach to source identification

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 contamination distributed across a network can best be solved by considering this network structure and the dimensions of information it contains. The approach proposed and developed in this thesis is built upon the fundamental concept that we can utilize this network structure and its multiple dimensions of information better solve the problem of tracing the source of large-scale outbreaks.

While the exact parameters will vary from product to product, all food distribution systems can ultimately be represented by the same layered, directed network structure (Figure 1). The distribution network is made up of multiple stages of production, distribution, storage, and consumption, where each stage represents a specific class of supply chain actors. Food is created in the first stage, which represents the point of production at a farm or other type of producer, and is distributed along links by logistic service providers (represented by links) to interior stage nodes until it reaches the final stage, representing point of sale at retail or food service establishment. The interior stages may be involved with storage, collection, or further processing of the commodity. The network in Figure 1 is composed of 17 producers, 7 processors, 7 distributors, and 21 retailers. The source of a large-scale, multi-state outbreak will originate at a high stage of the network, e.g. in the producing or processing stages; only nodes in these stages are able to reach downstream nodes in geographically distributed locations. Case reports of illness are associated with the retail node at which the offending product was purchased.

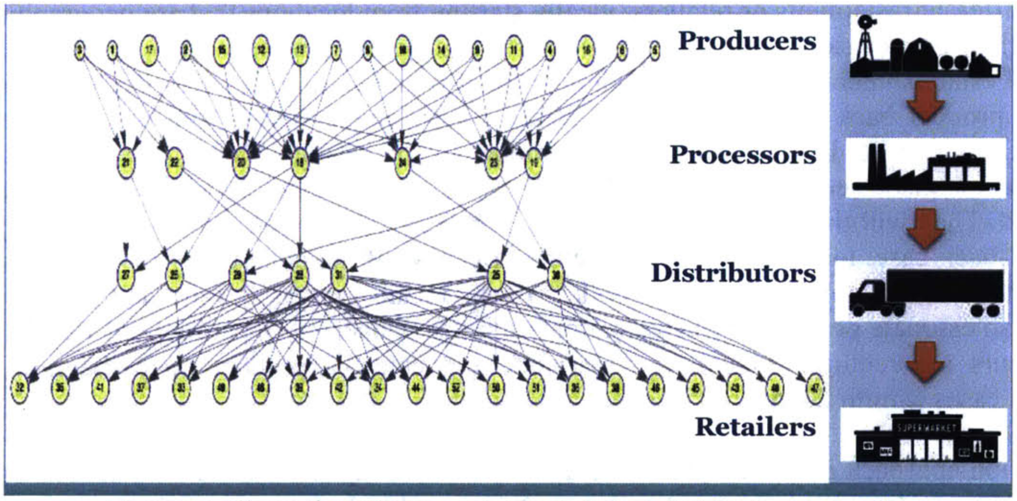


Figure 1. Illustration of a layered, directed food distribution network.

With this structure fully 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. The feasible sources can thus be identified as the set of nodes in the producing or processing stage that share at least one network path to all contaminated retailer nodes. The set of feasible sources resulting from utilizing all case report evidence will be smaller than that resulting from the subset considered in triangulation.

Network structural information provides a first cut into the source identification problem, enabling us to identify the feasible sources of contamination. To differentiate between the feasible sources, we can leverage further dimensions of information available *within* the network. First, each link contains information about the quantity or volume of goods traded between supply chain actors. Second, time dynamics along the links provide additional information. Each network path, or collection of directed links and nodes from origination to point of sale, contains information about the distribution of time that a contaminated product could have taken to travel these steps. Combining information from the volume along links and the resulting temporal distribution along paths provides insight into who is more likely to have transmitted to whom. This insight can help us to discriminate among the feasible sources.

How is this information being leveraged by existing approaches in the research literature? Most work involving networks and outbreaks of contamination has been focused on the forward problem of understanding and forecasting the spreading process and its dependence on the structure of the underlying network. A portion of this research has been specific to the case of foodborne disease outbreaks on food distribution networks. In recent years, we have seen some work on the inverse problem of identifying the source of general outbreaks. Nonetheless, none of these studies have been specific to the case of foodborne disease. The network structure assumed in all existing work is a general, undirected network in which any node can be the source and any node can be contaminated, which is not the case for food distribution networks. Furthermore, the approaches can be categorized as utilizing either network structure and volume information or network structure and temporal information, but not a combination of the two.

This thesis proposes that we can do better by tailoring an approach to the specific layered network structure of food distribution networks that leverages all dimensions of information available for solving this problem.

Our approach: Bayesian, network-theoretic source identification

In this thesis we develop a traceback approach built on the following core principles:

- *Utilize the network structure* – Design an approach specific to the directed, layered structure of the food distribution network
- *Utilize all available information* – Use information from all case reports, network trade flow volumes, and the temporal distribution of goods along network paths
- *Incorporate prior information* – Incorporate information external to the network, such as known risk factors and expert opinions if available

Based on these core principles, a source identification methodology has been developed that (1) immediately identifies all *feasible* source locations, (2) ranks the *feasible* locations by the likelihood that each one is the *true* source, (3) uses the ordered ranking to create systematic investigation strategies. When evaluated against existing methods in traceback, both those currently in practice and those proposed in academic literature, this methodology demonstrates significant benefits in accuracy, certainty and speed. Furthermore, these benefits are possible at low financial and opportunity cost to implement: low financial cost because the traceback system would function on a computer model at very low cost to implement and no cost to operate; low opportunity cost because generation of investigation strategies come at no exclusion of existing approaches.

Network traceback methodology

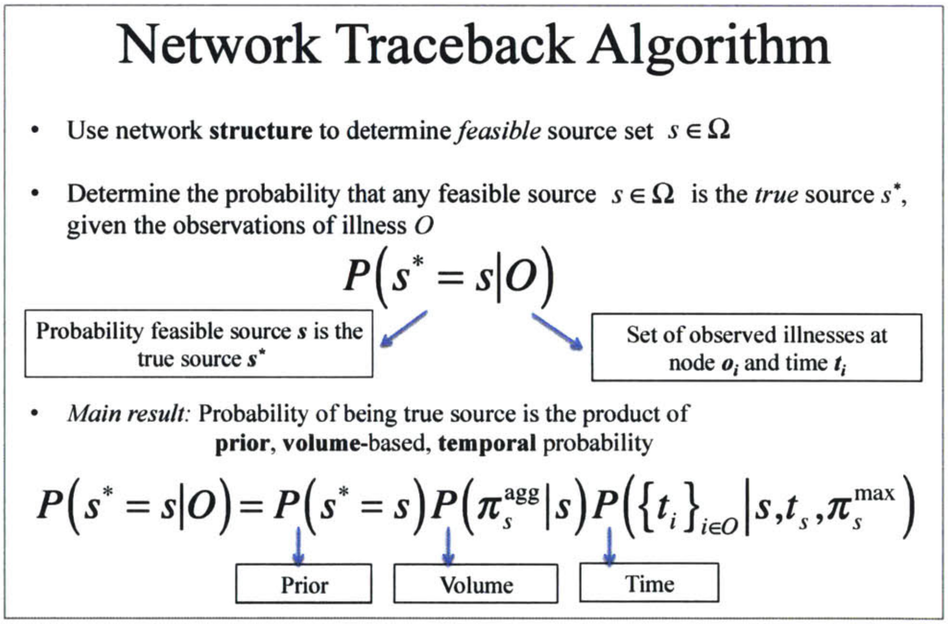
A source identification algorithm was designed to accomplish steps (1) and (2) above. The source localization algorithm requires the following input data:

- Food supply chain network information
    - *Identity* of supply chain nodes and *location* in geography

    - The *existence* of trade links between supply chain nodes and *volume* traded
    - *Time dynamics* of how contamination spreads across the network
- Case report data:
    - *Location* in the network
    - *Time* of occurrence, according to patient's recalled time of illness onset

The algorithm first performs a "preprocessing" step, using the network structure to determine the feasible sources as the set of processing or producing stage nodes that share at least one network path to all contaminated nodes. For any given instantiation of the algorithm, the source must be assumed to be in either the processing or producing stage.

With the feasible set identified, the algorithm then determines the probability that each feasible source is the true source, given the observations of illness at specific node locations and times. To determine this probability, a derivation is provided in the thesis that factors out the volume-based probability contribution, decomposing the probability of being the true source into a volume component and a temporal component. The result is that the *posterior* probability of being the true source is the product of the Bayesian prior probability, a volume-based probability factor, and a temporal probability factor (Figure 2). An approach to efficiently estimate the volume and temporal probability factors, accounting for the computational constraints of operations on networks, is designed in the thesis.



Network Traceback Algorithm

- Use network **structure** to determine *feasible* source set $s \in \Omega$
- Determine the probability that any feasible source $s \in \Omega$ is the *true* source s^*, given the observations of illness O

$$P\left(s^* = s \middle| O\right)$$

Probability feasible source ***s*** is the true source ***s****

Set of observed illnesses at node $\boldsymbol{o_i}$ and time $\boldsymbol{t_i}$

- *Main result:* Probability of being true source is the product of **prior**, **volume**-based, **temporal** probability

$$P\left(s^* = s \middle| O\right) = \underbrace{P\left(s^* = s\right)}_{\text{Prior}} \underbrace{P\left(\pi_s^{\text{agg}} \middle| s\right)}_{\text{Volume}} \underbrace{P\left(\left\{t_i\right\}_{i \in O} \middle| s, t_s, \pi_s^{\max}\right)}_{\text{Time}}$$

Figure 2. Overview of the Network Traceback Algorithm developed in this thesis.

The methodology determines the probability factors for **prior**, **volume**, and **time**, applying the same process for each feasible source node individually. As mentioned above, the **prior** probability is informed by information external to the network structure. If information on known risk factors or expert opinion is not available, the relative production quantity at each feasible node is used, assuming that any product produced is equally likely to generate contamination *a priori*. The **volume** contribution quantifies the probability that a feasible source node could have reached all contaminated nodes. Here, we are essentially assuming that the (relative) total volume of goods flowing from the source to all contaminated nodes is as a proxy for this probability. We calculate this term using the weighted adjacency matrix representing the network. The temporal contribution quantifies the probability that the feasible source generated the observed illness times, given what we know about the time dynamics from contamination origination to observation. In other words, this is the probability that the feasible source node can "explain" all of the observed contamination times. To determine this probability we first identify the set of highest probability paths from the feasible source to each contaminated node. We then find the start time that maximizes the likelihood of the observation times – and record the associated probability, which is the area under the likelihood curve within a small "uncertainty window" around the observed contamination times.

Once each probability factor has been determined for a feasible source node, we multiply these probabilities together to determine the posterior probability that node is the true source. We do this for each feasible source node, normalizing the resulting set to define a posterior probability mass function (PMF) representing the probability-ordered ranking over the set of feasible sources. An illustrative PMF and ordered ranking is depicted in Figure 3.

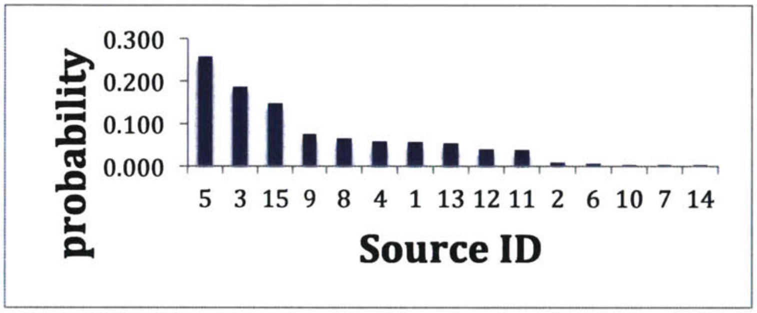


Figure 3. Illustrative Probability Mass Function (PMF) and ordered ranking of feasible sources resulting from applying the Network Traceback Algorithm to traceback an outbreak of foodborne disease.

Source identification methodology performance

We have subjected the methodology to an extensive performance evaluation. The primary goals of this performance evaluation study were to (i) establish the accuracy of the methodology, (ii) determine how the accuracy compares with existing traceback approaches, and (iii) ensure that these results are robust across many different network structures. A secondary objective was to explore the relationship of traceback accuracy on specific network parameters.

A Monte Carlo (MC) simulation environment was developed to measure traceback performance across a wide range of network structures and outbreak scenarios. Underlying this framework is a set of network structural models and contamination simulation model. We developed two types a network generating models: a "stylized" model and a geographically accurate model. The stylized model allowed us to generate networks according to a set of 10 variables determining the food distribution structural, volume, and temporal parameters. The regionally accurate generating model was implemented to produce networks representing the supply structure for tomatoes and lettuce grown in the US (Figure 4), informed by data from a combination of sources including the USDA, publically available retailer data, and the research literature. These realistic models represent various features of the real distribution system, featuring in particular the extreme clustering of production or cultivation in specific regional areas. As indicated in the figures, the model represents 13 clusters of tomato production across the country – this accounts for 80% of US tomato consumption, while lettuce production is even more aggregated – our network model represents 98% of US consumption by featuring only 3 growing locations.

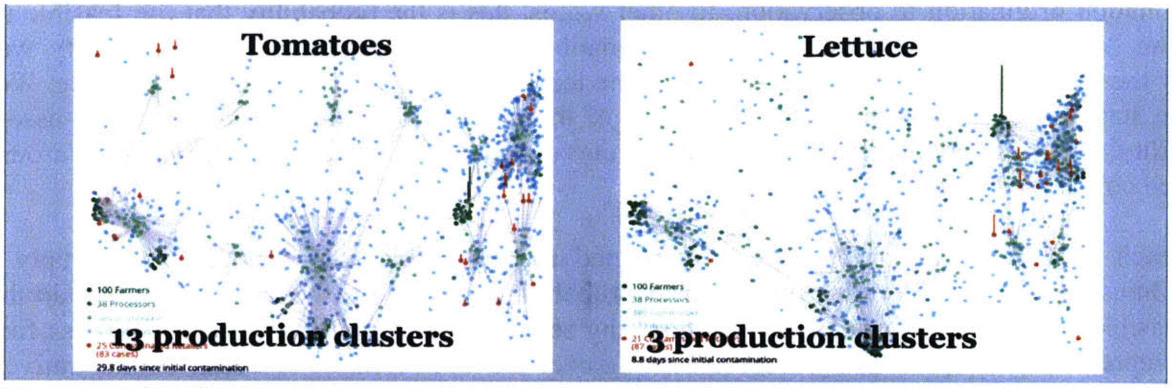


Figure 4. Visualization of an outbreak occurring in the tomato and lettuce network models. Supply chain nodes are represented in green. Locations reporting contamination are represented in red, with vertical red lines signifying the density of contamination reports at that location.

For any given network structure, the MC simulation model was used to generate outbreaks, trajectories of contamination through the supply chain, and reports of illness at specific times and node locations. At a slice in time in the outbreak's progression, the traceback algorithm was applied and a PMF over the feasible sources is constructed. The feasible sources are then rank-ordered according to their probability values. This process was repeated at various intervals as the contamination event progresses and illnesses continue to present to generate a series of rankings as a function of time or case development. To assess the traceback performance for a particular network structure, multiple contamination events were generated and traced back, and the cumulative results assessed according to a set of accuracy metrics including:

- *Traceback Accuracy*, the percentage of times the true source is accurately identified
- *Rank of True Source*, the position of true source within the ordered ranking
- *Distance from True Source*, the geographical distance between the top ranked source and the true source

The computational results presented in the thesis serve as a powerful first step towards validation of the accuracy and applicability of our approach. For stylized and realistic networks, the method performs well and follows expected properties, increasing with data on the number of contamination reports. We find that we can make very good inferences about the source location after only a limited number of illnesses have been reported, and very accurate inferences if we wait a bit longer. These conclusions apply both to locating the source location at a network node and on the map, with *traceback accuracy* ranging between 80 – 95% and the *distance from the true source* ranging from 5 – 15 miles, for the specific networks considered. These performance results suggest that our traceback approach provides an effective framework for identifying the source of large-scale outbreaks of foodborne disease.

We go on to demonstrate the benefits of our approach in comparison with existing approaches to traceback. We implement the *FDA Heuristic,* which models the existing regulatory traceback approach of triangulation; the *Network Baseline*, which uses network structure and the full set of case reports to identify feasible sources but does not distinguish between these sources probabilistically; and a best-in-class method presented in the literature (Brockmann and Helbing 2013) from the category of approaches using network structure and volume (but not temporal) information. We compare these approaches to our method, which combines network structural, prior, volume, and temporal information to distinguish between sources. In results from tracing back the source of multiple simulated outbreaks generated across a large set of network structures, we consistently observe that the following set of relationships hold:

- Considering the network structure and the full available set of case data improves upon the narrow view afforded by the triangulation method currently applied in investigations.

- Using what is known about the volume of food flows through the distribution structure to differentiate between feasible sources substantially improves upon the resulting accuracy.
- Designing the approach specifically for the layered structure of food distribution networks and incorporating prior, volume, and temporal information further improves upon the volume only, general network approach.

This series of increasing relationships is visible in the two examples depicted in Figure 5. Each example illustrates *traceback accuracy* as a function of the number of reported cases of illness for a different stylized network structure. Compared with the current regulatory approach of triangulation (light green), the improvement in accuracy with our approach (blue) is always significant, exceeding 80% accuracy in all cases evaluated. This improvement in accuracy suggests that our network-theoretical approach to source identification can contribute substantially to the existing traceback investigation process.

Because the traceback results are dependent on the parameters of the network structure, the magnitude of the jumps in accuracy will vary. As can be seen in the Figure 5 examples, when compared to the existing approach in the literature (in red) the improvement in accuracy possible with our methodology can range from extreme (85% with stylized network – high variance) to more moderate (10% with stylized network – zero variance), but is observable in all cases evaluated. This improvement demonstrates the theoretical value of designing a traceback methodology specific to the problem of foodborne disease outbreaks on food distribution networks.

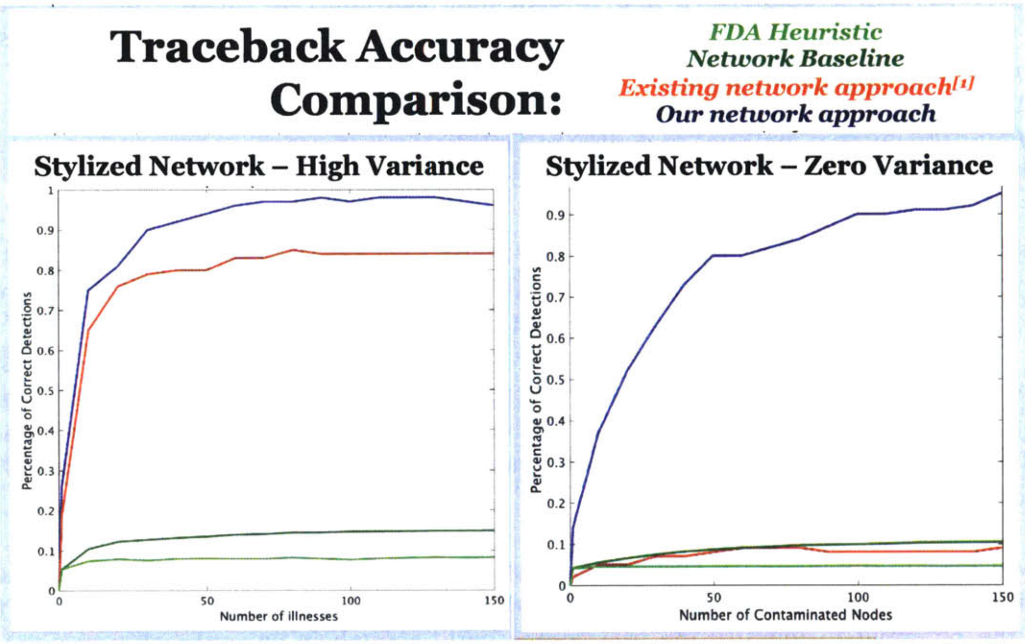


Figure 5. Performance of our approach to traceback (in blue) compared with existing approaches: the FDA Heuristic (light green), the Network Baseline (dark green), and a network structure and volume utilization approach from the research literature (red). Each figure presents the *traceback accuracy* as a function of the number of reported cases of illness for a different stylized network structure.

Implications for traceback response

The performance evaluation suggests that our traceback approach may provide substantial benefits to investigators during traceback investigations. The next step is to develop a decision-making framework to guide investigators at the tactical level to make the most effective interventions to solve an investigation and stem impact on the public.

We can use the cumulative output from multiple traceback results from simulation to quantify the *variability* of detection performance. The variability can be understood as the range of positions taken by the true source within the resulting ordered ranking. For example, considering the tomato network and applying traceback after the first 10 reports of illness, the true source is identified in first ranked position in ~70% of simulations, within the first 2 positions in ~80% of simulations, and within the first 5 positions in 95% of results. Quantifying variability thus allows us to say that we can expect to identify the true source within a specific number of high probability (or top-ranked) candidates with a specific level of accuracy. For practical purposes, the size of this bounded set effectively quantifies *the number of source candidates necessary to investigate to identify the true source*.

The size of the bounded set also quantifies the *improvement over time*. As the number of case reports increase and the accuracy improves, the variability decreases. Accordingly, the number of source candidates necessary to investigate to identify the true source will be reduced. Continuing with the tomato network example above but applying traceback after 30 reports of illness, the true source is identified in first position in 90% of simulations and within the top 2 positions in 95% of simulations. This reveals an important tradeoff for investigators: wait for a certain number of illnesses to accrue until the source can be uniquely identified with very high accuracy, or act early to prevent further illnesses but at a greater cost.

These insights inform decision making strategies for investigation interventions, including when and to which facilities to send investigators, and when and which facilities to target in public service messaging. In the thesis we develop a framework that investigators can apply to identify and compare various investigation intervention strategies. This framework is based on a set of performance attributes:

- *Accuracy*, the precision in correctly identifying the source within a specific number of top-ranked predictions
- *Specificity*, the number facilities to deploy to / implicate in message
- *Benefit to Public Health*, the number of illnesses potentially averted

These performance attributes allow investigators to specify their desired accuracy level (i.e. the level of risk they are willing to take on) and available budget, then to determine the expected number of illnesses that can be averted for these values. Conversely, the attributes allow quantifying the cost of reducing more illnesses. For example, in computational results demonstrated for the tomato network discussed above, we observe that with 2 investigators and 95% accuracy required, 42% of the total cases resulting from the outbreak can be averted. With 3 investigators and the same accuracy requirement, the investigation can be launched earlier and 50% of cases can be averted, and when 4 investigators are available, 60% of cases can be averted.

Beyond demonstrating the value a systematic framework for investigation response, these results highlight a more fundamental takeaway: that the traceback methodology developed this thesis demonstrates the potential to identify the true source early enough in an outbreak's progression that a substantial fraction of the illnesses can be averted. While the actual number of illnesses that can be averted will depend on the particular outbreak scenario, the fact that cases can be averted is a huge result in itself.

Conclusions and next steps

In summary, this thesis develops a novel network approach to traceback of foodborne disease, a novel traceback methodology that is specific to the case of food distribution networks and foodborne disease, and a recommended approach for investigators or emergency responders to act on this information. Computational results suggest this set of methodologies can contribute major improvements to outbreak response on three important dimensions:

1. ***Outperform current triangulation methods***: The true source was identified with >80% greater accuracy
2. ***Successfully resolve many more investigations***: The true source was identified with 95% accuracy using the evidence from only 10 – 30 cases of illness
3. ***Resolve investigations early enough that cases can be averted***: 40 – 60% of illnesses averted in simulated interventions

These computational results are very promising. However it is important to stress that these benefits are estimated on the basis of results of simulation; live use of these techniques may demonstrate features of the real problem not incorporated into these research and modeling efforts. There are multiple challenges to real-time implementation of the methodology including the delay in case reporting and the underreporting of cases; the uncertainty in network distribution times; and the imprecision of patient-recalled time of illness onset. There is the additional fundamental challenge of constructing a database of network models for various foods so that the methodology is ready to be deployed in real-time in the event of an outbreak. The implementability of the traceback methodology ultimately depends on access to representative network models. However, acquiring and organizing this information presents three major challenges. First, food distribution networks are markets characterized by inherent stochasticity, which can be challenging to model. Second, network models require aggregated food distribution data that is not readily available for public use or not systematically recorded. Third, collection and organization of *available* data present extensive practical data-management challenges. Foods of today are complex and outbreaks can occur in foods containing dozens of ingredients. Hugely complex trade network would result from the consideration of all supply chain actors, big and small, and characterizing all commodity flows as well as external trade relations with different producers in the industry.

There is strong potential that a recently developed food distribution network model developed by researchers at Kühne Logistics University (KLU) and the Technical University of Darmstadt in Germany can be used to overcome these challenges (Friedrich 2010, Balster and Friedrich 2016). Their model utilizes only existing, readily available data sources coming from public authorities, food-related associations, and professional data providers. It covers the supply of 50 different foods across Germany's food industry, accounting for interactions between these foods, making it extendable to tracing processed foods containing dozens of individual ingredients.

The next step will be to deploy our traceback methodology with the KLU food distribution network model. We will work together with researchers at the KLU to integrate their network model with our traceback methodology to form a *holistic traceback framework*. Using this combined model and method approach, we will seek to demonstrate the ability to identify the origin of recent outbreaks that have occurred in Germany. Success in correctly identifying the source of these outbreaks will be an important step in validating the accuracy and effectiveness of our methodology for identifying the source of large-scale outbreaks of foodborne disease.

Chapter 1: Introduction

Foodborne illness outbreaks impose enormous health and economic burdens in the US. Identifying the spatial origin of the contaminated food causing an outbreak is a challenging problem due to the complexity of the food supply and the absence of coherent labeling and distribution records. Current investigative methods are slow, resource intensive, and the overwhelming majority of investigations are unsuccessful in identifying the location source of an outbreak. New tools and approaches that take advantage of modern data and analytical techniques are needed to more quickly identify outbreak origins and prioritize response efforts.

The practical objective of this thesis is to contribute to the ability to efficiently locate the source of large-scale outbreak while contamination-caused illnesses are occurring, thereby resolving investigations earlier and averting potential illnesses. To this effect, our primary contribution is the development of a holistic system for rapid identification of the source within the constraints of available or acquirable data and resources. The system is based on a novel, network-theoretic approach for outbreak detection that functions on current domestic food systems network data, reported cases of illness and a prior probability function over likely sources. The traceback framework outputs a set of feasible sources of contamination and their relative probability of being the *true* source, producing an ordered ranking over the feasible sources. This source data forms the basis of a decision model for investigation interventions, which provides investigators with a set of recommendations for two distinct options for action: (i) where to send on-the-ground investigators and (ii) when and what to message to the general public about potential contaminated foods. The development and review of this holistic framework is provided in the four main chapters of this thesis and a future plan for practical validation is provided in the conclusion.

In this chapter, we discuss the problem of tracing back large-scale, multi-state outbreaks of foodborne disease. In Section 1.1, we describe the growing impact and prevalence of these outbreaks. In Section 1.2, we overview current regulatory approaches to outbreak investigation and response, highlighting the major sources of delay and opportunities for improvement. In Section 1.3, we describe existing interventions and efforts to improve foodborne disease investigations and identify outbreak origins through development of technologies and legislation. We then review approaches to contribute to this problem in the existing literature, namely (i) risk-based and (ii) tactical approaches to traceback. In Section 1.4, we present the main contribution of this thesis: the development of novel, integrated system for rapid identification of the source.

1.1. The Growing Impact of Large-Scale, Multi-State Outbreaks of Foodborne Disease

The complexity and globalization of food production have made foodborne disease a widespread public health problem in both developed and developing countries. The Center for Disease Control (CDC) estimates that each year in the U.S. 48 million illnesses, 128,000 hospitalizations, and 3000 deaths result from foodborne disease, at an estimated annual loss of $152 billion (Osterholm 2011). Due to underreporting or unknown agents, these numbers are potentially much higher (Batz 2005, Batz and Morris 2012, Scallan et al. 2011).

Most foodborne outbreaks involve small groups of people in a localized area. Although the probability of any given food generating a large, multi-state outbreak is extremely low, low probability events do happen and may have massive impact. In the summer of 2011, an outbreak of E. coli O104:H4 from sprouts grown in Germany caused 55 deaths and 4,075 illnesses in 16 countries in the five-week period it took investigators to identify the source of contamination (WHO 2011). Earlier that year, an outbreak of listeria linked to cantaloupes grown in Colorado ravaged across the country. Over a timeline of almost three months, it took 30 lives and infected 146 people across 28 states (CDC 2011). As recent trends continue, including large-scale production practices and distribution over ever-larger distances, these major outbreaks are increasing in both prevalence and in the severity of consequences. From 2005 – 2014, nearly 200 multi-state outbreaks were identified and investigated in the US as compared with 85 over the years 1995 – 2004; these multi-state outbreaks accounted for 3% of total outbreaks, but were responsible for 34% of hospitalizations and 56% of deaths (Crowe et al. 2015).

Furthermore, this streamlining in the mass production of food also represents vulnerabilities to attacks by deliberate contamination, i.e. biological or chemical terrorism. While cases of intentional contamination have been infrequent, there is a growing concern among the U.S. intelligence agencies that these vulnerabilities, together with changes to the threat environment, have made agroterrorism a more viable and attractive approach for adversaries (Decker 2014, FDA 2013a, Olson 2012).

Recognizing when outbreaks occur, moving swiftly to respond, and obtaining information to prevent future outbreaks are critical parts of maintaining a safe food supply. Current research efforts and federal initiatives to mitigate the impact of foodborne disease outbreaks, including the landmark 2012 Food Safety Modernization Act (FSMA), have focused largely on addressing the causes of outbreaks (Nuzzo 2013). While these efforts are essential to minimizing the risk of illness associated with the consumption of contaminated produce, they do not provide the tactical support necessary for response to foodborne illness outbreaks that occur when such preventive efforts have failed (IOM 2010).

When people begin to fall ill during an outbreak, time is of the essence. Yet complex market behaviors including uncertainty in sourcing and complicated distribution chains make it especially difficult to establish provenance of our food supply. The complexity, dynamics, and massive size of food supply chains means there are a huge number of potential sources of outbreaks and it is not feasible to test them all, or even a meaningful fraction of them, in a short period of time. As an example, consider the source of the seven main ingredients that compose a single loaf of Sara Lee Bread. Not even considering foreign farms and processing sites, these ingredients can originate from 2.1 million farms, pass through 30,000 processing facilities and 19,000 (re)packers, to end up at over 1.1 million retail food stores and outlets, as illustrated in a New York Times article in 2007 (Schoenfeld 2007). In addition, delays and inefficiencies in the foodborne disease investigation process limit our national ability to identify the origin of foodborne illness outbreaks in a timely manner. Current practices are time consuming, resource intensive, and outdated in methods that do not take advantage of modern data and analytical techniques or methods.

As a result, despite the dedicated efforts of food safety officials across the country, our current capacity to identify the origin of illness in multi-state outbreaks is characterized by typical timelines of 1-2 months, completed in many cases after the outbreak is almost over – and these are the minority of outbreaks for which the traceback is successful (Wilkins et al. 2015, McEntire and Bhatt 2013). In the 13,352 foodborne disease outbreaks (causing 271,974 illnesses) documented by the CDC during 1998-2008, only 4,887 (37%) were traced back to a single food vehicle and pathogenic source, with less than 15% of these to a specific contamination point (Painter et al. 2013).

1.2. Background on Foodborne Disease Outbreaks and Traceback Investigations

A foodborne disease outbreak investigation begins from the time the first case presents symptoms and ends when the contaminated product has been conclusively identified. The investigation is a multi-disciplinary task that requires information from many sources including laboratory work, patient interviews, environmental and food preparation reviews, and collection of distribution records or traceability data, when available.

There are three standard components to the outbreak response and investigation process, beginning from the time the first case presents symptoms and ending when the contaminated product has been conclusively identified: (i) detecting that an outbreak is occurring, (ii) identifying the food vector causing the outbreak, and (iii) identifying the location source of the outbreak at a farm or processing center. There is no general procedure for foodborne disease outbreak response that fits every event perfectly, and these three components are not necessarily conducted sequentially. To detect the outbreak, an unusual accumulation of disease reports has to be identified and confirmed as pertaining to the same strain. This involves laboratory tests to specify the pathogen and corresponding microbiological "fingerprinting" or genetic sequencing of strains to verify case relatedness. After the outbreak has been confirmed, the investigation to identify the transmission vehicle and location source of contamination begins. The vehicle is identified through a standard epidemiological process of interviewing cases to identify common foods consumed, combining this with microbiological sampling of culprit foods and food surfaces. After the food source is identified, or concurrently with that investigation, an analysis is undertaken to determine the location origin of the outbreak. The current regulatory approach generally involves "triangulation," or tracing back the distribution paths of products from several locations to determine if there is a common point of convergence in the supply chain, for example a common date and location of harvest or place of manufacture. If identified, the contaminated foods can be traced forward from the contamination origin to determine transmission routes and identify products and consumers at risk. A final stage, which may occur well after any preventive action can be taken to limit the number of illnesses in some investigations, is the evaluation of the specific practices at the farm, transportation, or other facility that may have caused outbreak. (FDA 2001, Wilkins et al. 2015, WHO 2008).

Tracing the source of foodborne illnesses is very complicated, especially for fresh produce items that have no bar codes, no packages, and are quickly consumed, often with other produce. The challenges of a fresh produce traceback investigation include the complexity of distribution systems and multiple sources of product at the point of sale, underreporting and significant time lags in the effects of contaminated food, and inconsistencies and gaps in labeling and distribution records. In addition to these many inherent challenges to the traceback problem, limitations to the foodborne disease investigation process make it difficult to improve our national capacity for detecting the origins of foodborne illness in a timely manner. We now provide a more detailed look at the essential elements of the investigation process, highlighting the major sources of delay and opportunities for improvement.

Identifying an Outbreak

Traceback investigations are not initiated until a critical cluster of cases have been identified and linked to a single outbreak strain by PulseNet, the national network of public health and food regulatory agency laboratories coordinated by the CDC that perform standardized molecular subtyping ("DNA fingerprinting") of foodborne disease-causing bacteria to distinguish strains at the DNA level (CDC 2015a). The second line of Figure 1.1 (red boxes) represents the case reporting timeline, which runs from the time a patient begins to experience symptoms of illness after eating a contaminated food, to the pursuit of medical attention, the laboratory tests to diagnose the causative agent, and the final confirmation through PulseNet that the case is part of an outbreak. The timeline for case reporting is presented in further detail in Appendix 1.1. This timeline can range from days to months, and means there can be a substantial delay between the start of illness and confirmation that a patient is part of an outbreak.

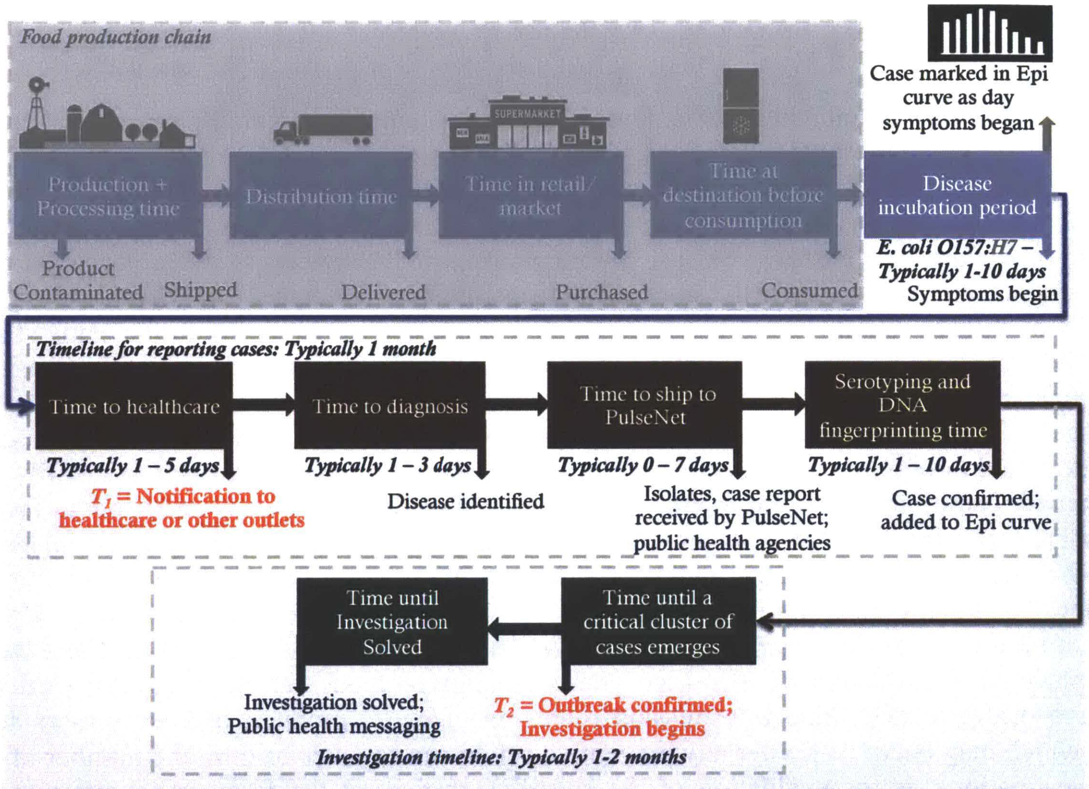

Food production chain
Production + Processing time
Distribution time
Time in retail / market
Time at destination before consumption
Disease incubation period
Product Contaminated
Shipped
Delivered
Purchased
Consumed
Case marked in Epi curve as day symptoms began
E. coli O157:H7 – Typically 1-10 days
Symptoms begin
Timeline for reporting cases: Typically 1 month
Time to healthcare
Time to diagnosis
Time to ship to PulseNet
Serotyping and DNA fingerprinting time
Typically 1 – 5 days
Typically 1 – 3 days
Typically 0 – 7 days
Typically 1 – 10 days
T_1 = Notification to healthcare or other outlets
Disease identified
Isolates, case report received by PulseNet; public health agencies
Case confirmed; added to Epi curve
Time until Investigation Solved
Time until a critical cluster of cases emerges
Investigation solved; Public health messaging
T_2 = Outbreak confirmed; Investigation begins
Investigation timeline: Typically 1-2 months

Figure 1.1. Graphic representation of timeline of foodborne disease contamination dispersion, case reporting (characteristic of E. coli O157:H7), and investigation. Data sources: (CDC 2015a-b, FDA 2001, Wilkins et al. 2015, WHO 2008).

A "critical cluster" is defined by the CDC to be a larger number of people having the same illness in a given time period than expected from long-term surveillance (CDC 2015d). While no statistical measure is used here, the number of cases that is large enough to distinguish the set as 'critical' will depend on multiple variables, including the nature of the outbreak – its size and dispersion, the virulence of the pathogen, and the time of year the outbreak occurs. This can be illustrated by considering a specific example, the 2013 outbreak of *Cyclospora* in fresh cilantro, analyzed in detail in Section 3. Table 1 depicts confirmed cyclosporiasis cases by week of illness onset for the 2013 outbreak compared to the long-term weekly mean for that time of the year (CDC 2013). The earliest this set of cases would start to look like an outbreak would be week $n = 1$ of the outbreak, the week of June 2, when the count of 12 cases is noticeably greater than the weekly average count of 6.8. Notably, June is the month of the year with the highest number of outbreak cases; if the outbreak occurred earlier in the year, fewer cases would be necessary to show a significant rise above the mean.

Week *n*	Week of Onset	Number of Cases	5-Year Weekly Mean
0	May 26–Jun 1	10	7.6
1	Jun 2–Jun 8	12	6.8
2	Jun 9–Jun 15	34	6.2
3	Jun 16–Jun 22	40	5.6
4	Jun 23–Jun 29	99	4.8
5	Jun 30–Jul 6	69	2.2
6	Jul 7–Jul 13	57	3.8
7	Jul 14–Jul 20	46	1.4
8	Jul 21–Jul 27	18	1.4
9	Jul 28–Aug 3	10	1.6
10	Aug 4–Aug 10	5	0.8
11	Aug 11–Aug 17	3	1.8
12	Aug 18–Aug 24	2	0.8

Table 1.1. Confirmed cyclosporiasis cases by week of illness onset compared to the 5-year weekly mean. Data source: (CDC 2013).

Because of delays in case reporting for a cluster of cases, and in particular the series of tests at both local diagnostic laboratories and PulseNet facilities, it is typically several weeks before health agencies confirm that an outbreak is occurring and react in the form of a traceback. While DNA fingerprinting is a vital step to establishing with certainty the link between nationwide cases of infection to a common outbreak, this convention poses a major limitation on the timeline of an outbreak investigation (Wilkins et al. 2015).

Identifying the outbreak source

Once the outbreak has been confirmed, the second and third stages of the investigation to identify the food and location source of contamination can begin. A typical large-scale nationwide foodborne disease investigation will first implicate the (set of) food product(s) responsible for the outbreak. As the very first step, an epidemiological investigation is conducted to implicate potential commodities by interviewing initial cases with regard to common factors and sampling food specimens as potential sources of contamination. A case definition is often established to identify further outbreak related cases and to collect information in a standardized survey. Using this data, analytical investigations, such as case-control and cohort studies, are performed to test hypotheses about the transmission vehicle. Ideally, a single food item is implicated and the second stage of the traceback is begun to trace the outbreak to its specific source location. In many cases, the combination of a small and slowly growing number of confirmed cases of illness, inaccuracies in individuals' recollections, and the common consumption of bundled foods (i.e. salsas, burritos, salads) make it impossible to statistically narrow down the search to a single item (McEntire and Bhatt 2013). In these situations, the two stages of the investigation may be conducted in parallel.

Following that identification, investigators will trace back the distribution of the implicated product and determine the source location (FDA 2001, Wilkins et al. 2015, WHO 2008). This stage of the investigation generally involves "triangulation," or tracing back the distribution paths of products from several locations to determine if there is a common point of convergence in the supply chain, for example a common date and location of harvest or place of manufacture (FDA 2001, Wilkins et al. 2015). To prioritize leads for the analysis, investigators consult with colleagues at state public health agencies, at university agricultural research institutions, and in industry to prioritize commodity sources. Notably, however, a systematic method for prioritizing sources does not currently exist (S. McGarry, personal communication, December 20, 2012).

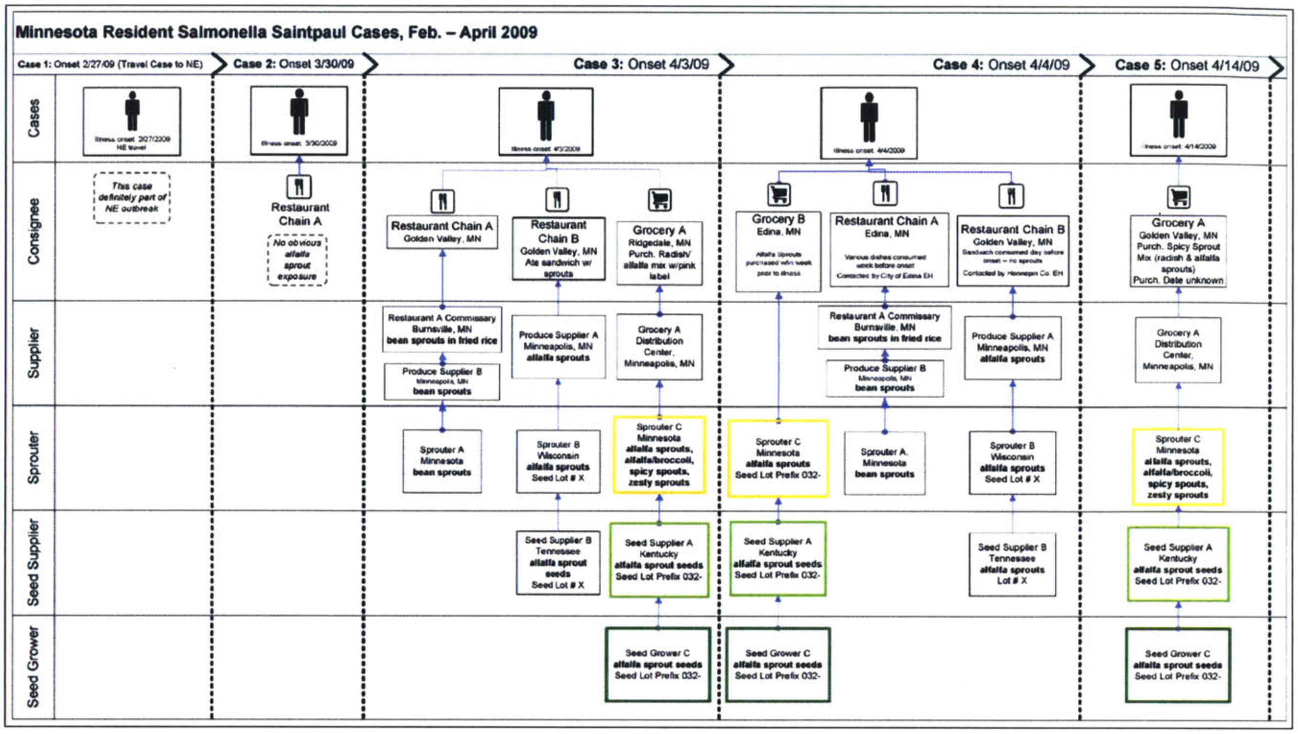

Minnesota Resident Salmonella Saintpaul Cases, Feb. – April 2009
Case 1: Onset 2/27/09 (Travel Case to NE)
Case 2: Onset 3/30/09
Case 3: Onset 4/3/09
Case 4: Onset 4/4/09
Case 5: Onset 4/14/09
Cases
Consignee
Supplier
Sprouter
Seed Supplier
Seed Grower
Illness onset 2/27/2009 NE travel
This case definitely part of NE outbreak
Illness onset 3/30/2009
Restaurant Chain A
No obvious alfalfa sprout exposure
Illness onset 4/3/2009
Restaurant Chain A Golden Valley, MN
Restaurant Chain B Golden Valley, MN Ate sandwich w/ sprouts
Grocery A Ridgedale, MN Purch. Radish/ alfalfa mix w/pink label
Restaurant A Commissary Burnsville, MN bean sprouts in fried rice
Produce Supplier B Minneapolis, MN bean sprouts
Produce Supplier A Minneapolis, MN alfalfa sprouts
Grocery A Distribution Center, Minneapolis, MN
Sprouter A Minnesota bean sprouts
Sprouter B Wisconsin alfalfa sprouts Seed Lot # X
Sprouter C Minnesota alfalfa sprouts, alfalfa/broccoli, spicy spouts, zesty sprouts
Seed Supplier B Tennessee alfalfa sprout seeds Seed Lot # X
Seed Supplier A Kentucky alfalfa sprout seeds Seed Lot Prefix 032-
Seed Grower C alfalfa sprout seeds Seed Lot Prefix 032-
Illness onset 4/4/2009
Grocery B Edina, MN Alfalfa Sprouts purchased w/in week prior to illness
Restaurant Chain A Edina, MN Various dishes consumed week before onset Contacted by City of Edina EH
Restaurant Chain B Golden Valley, MN Sandwich consumed day before onset – no sprouts Contacted by Hennepin Co. EH
Restaurant A Commissary Burnsville, MN bean sprouts in fried rice
Produce Supplier B Minneapolis, MN bean sprouts
Produce Supplier A Minneapolis, MN alfalfa sprouts
Sprouter C Minnesota alfalfa sprouts Seed Lot Prefix 032-
Sprouter A. Minnesota bean sprouts
Sprouter B Wisconsin alfalfa sprouts Seed Lot # X
Seed Supplier A Kentucky alfalfa sprout seeds Seed Lot Prefix 032-
Seed Supplier B Tennessee alfalfa sprouts Lot # X
Seed Grower C alfalfa sprout seeds Seed Lot Prefix 032-
Illness onset 4/14/2009
Grocery A Golden Valley, MN Purch. Spicy Sprout Mix (radish & alfalfa sprouts) Purch. Date unknown
Grocery A Distribution Center, Minneapolis, MN
Sprouter C Minnesota alfalfa sprouts, alfalfa/broccoli, spicy spouts, zesty sprouts
Seed Supplier A Kentucky alfalfa sprout seeds Seed Lot Prefix 032-
Seed Grower C alfalfa sprout seeds Seed Lot Prefix 032-

Figure 1.2. Example of a product trace diagram illustrating exposure distribution pathways documented during the traceback of an outbreak of salmonellosis associated with alfalfa sprout consumption conducted by the Minnesota Department of Agriculture (Smith et al. 2015). Convergence points are indicated by boxes outlined in the same color.

Triangulation requires inspection of common distribution sites, processors, or growers through interviews, observations, and record collections. An example of a product trace diagram illustrating exposure distribution pathways and convergence points documented during the traceback of an outbreak of salmonellosis associated with alfalfa sprout consumption conducted by the Minnesota Department of Agriculture is depicted in Figure 1.2 (Smith et al. 2015). Points of convergence are indicated by boxes outlined in the same color.

This data collection is both resource and time intensive. The food system is complex with many possible supply chain pathways leading to each chosen location, all of which must be traced independently along the supply chain. Furthermore, collecting this data along each step poses a delay due to the many inconsistences and gaps in recordkeeping data. While many producers, manufacturers and retailers have product tracing systems in place, these systems vary greatly depending on the amount of information recorded, how far forward or backwards in the supply chain the system tracks, technologies used to maintain records, and the precision with which a system can pinpoint a product's movement (Golan et al., 2004; Wu et al., 2011, Storoy et al., 2013). The only legal requirement concerning product traceability in food supply chains is so-called "one-up, one back" recordkeeping, mandated by the Bioterrorism Act of 2002. The 2011 Food Safety Modernization Act (FSMA) now requires some additional recordkeeping for high risk foods (the BT Act, H.R. 3448; FSMA 2011). Requirements based on the Bioterrorism Act include having firms know who they received products from and to whom they were sent ("one up, one back" tracing), however some supply chain members, such as restaurants and farms, are exempt. The level of detail and the specific types of information required to be maintained depend on the role of the firm in the supply chain. Furthermore, FDA is often only provided access to proprietary information on trade flows when responding to an emergency incident, and only then when there is "reasonable cause" to make inquiries into the traceback information for a particular company; thus, this information is not organized in advance of a contamination event.

The challenges of triangulation pose a major limitation on the speed of the investigation, and thus the time it will take to identify the location source. According to an estimate provided by investigators with the Minnesota Department of Public Health, 8 to 24 person hours are required to collect paperwork and create a product trace diagram for 1 to 2 contamination cases (Smith 2015). As a result, our current capacity to traceback the sources of illness in multi-state outbreaks is characterized by typical timelines of 1-2 months, completed in many cases after the outbreak is almost over – and as discussed above, these are the minority of outbreaks for which the traceback is successful (Wilkins et al. 2015, McEntire and Bhatt 2013).

1.3. Opportunities to improve outbreak investigations

If public health officials can more quickly and successfully recognize when a foodborne illness outbreak has occurred and identify the food and location source of contamination, lives can be saved and economic losses averted. To achieve these goals, new tools, technologies, and the availability of digital sources of data are facilitating the development of novel strategies to contribute to the three stages of the outbreak investigation process. In this section, we overview existing efforts to improve traceback investigations. In particular, we highlight the need for improving efforts to contribute to the ability to *identify the location source* of an outbreak. Given this need, we propose a novel, network-theoretic approach to traceback. Developing this approach is the focus of this thesis.

1.3.1. Existing Efforts to Improve Foodborne Disease Outbreak Investigations

Detecting an outbreak using Next Generating Sequencing and Digital Disease Detection

The time until a common outbreak strain is identified and thus an outbreak is detected is rapidly decreasing as the use of whole genome sequencing technologies, commonly referred to as "next generation sequencing" (NGS), become pervasive. These technologies can sequence in one step and almost real-time what it takes current strain subtyping methods (i.e. "DNA fingerprinting") two to three laboratory tests conducted typically over several weeks perform with less precision. NGS technologies have been developing over the past decade but it is only now that they are becoming affordable enough to introduce at scale by the CDC-directed national network of state laboratories.

Digital Disease Detection (DDD) methods are being introduced to gather additional, publically-available data that can be applied to further decrease the time to detect an outbreak. While foodborne illness is notoriously under-reported, online illness reports have been shown to supplement traditional surveillance systems in detecting individual cases linked to known outbreaks or otherwise undocumented outbreaks by capturing reports from those who do not contact a health department (Nsoesie et al. 2014, Harris et al. 2014, Harrison et al. 2014). For this purpose, consumer self-reported concerns regarding foodborne illness, implicated foods, and consumption location can be crowdsourced from various popular social networking sites (e.g. Twitter, Facebook, GrubHub, Foursquare, and Yelp). Furthermore, dedicated resources such the website IWasPoisoned.com allow consumers to directly communicate their adverse food safety events on the public domain, information which is increasingly being picked up by public health departments and other food safety regulators. The real time identification of a greater sample of cases can accelerate the traceback timeline by contributing to identifying outbreaks earlier (as well as to improving the source localization process, as we will describe in later chapters). The development of surveillance tools that aggregate online information to contribute to these goals face challenges pertaining to inaccurate or noisy information, required participation levels, and privacy concerns. Still, these methods demonstrate great promise in accelerating the outbreak detection process.

Identifying the food source using retail sales data

Methods are being developed to use readily available data to supplement the existing regulatory approach to identifying the food vector carrying a contamination. Doerr et al. (2012) introduce a likelihood-based method to compare the distribution of sales data for a dataset of finished food products with the distribution of public health case reports in order to determine the product most likely to be associated with a foodborne disease outbreak. The approach leverages readily available product sales data already collected as part of routine business practices by retailers and distributors. The likelihood approach is evaluated in Kaufman et al. (2014), demonstrating high performance accuracy in identifying the culprit food product across a large set of simulated contamination scenarios. These methods are ultimately limited by the extensiveness and temporal relevance of the available food product dataset. However the success of simulation results demonstrated in Kaufman et al. (2014) suggest that the approach can provide public health investigators with better information than currently exists on which food products might be causing an outbreak.

Identifying the location source using technology enabled traceability systems

The use of technology-enabled systems to follow the movement of products through the supply chain can facilitate immediate traceback capabilities (Business Insights 2010, Wu et al. 2011, Storoy et al. 2013, McEntire and Bhatt 2013). A software enabled full-chain traceability system can capture the movement of each parcel or piece of food during its entire journey through the supply chain using electronic tags in the form of a barcode or radio frequency identification (RFID) chip. This information can be used to track and identify the real-time location of a food product no matter the complexity of the supply chain (Wu et al. 2011).

However even as the industry gradually moves towards a sensor-based future, full supply chain traceability, which requires all members of the supply chain to participate and pay, is a distant reality. Furthermore, there are limits to the power of technology-enabled traceability systems. Many food products have yet to become technologically traceable, and are unlikely candidates due to cost considerations or the nature of the product (Golan et al. 2004). And even if traceability data is systematized by individual businesses, tracking products across companies would still be the primary challenge. Golan et al. point out that while firms have an incentive to create systems that identify and isolate unsafe foods and remove them from their own supply chains, they are not incentivized to create a traceability system that tracks food beyond those borders. Furthermore, many firms find value in some level of anonymity. The desire for anonymity can be explained in that traceability systems increase the probability that a firm will be identified and exposed to liability in the case of food safety problems (Golan et al. 2004, Pouliot and Sumner 2008). Since the FDA is only provided access to proprietary information on trade flows when responding to an incident, and only then when there is "reasonable cause" to make inquiries into the traceback information for a particular company, this information cannot be organized in advance of a contamination event. Connecting the dots between companies can require considerable time and resources, delaying the traceback process (McEntire and Bhatt 2013).

1.3.2. Opportunities for improvement

This thesis focuses on improving the response to large-scale, multi-state outbreaks. There are multiple opportunities for improving each component of the outbreak investigation process that can have positive and meaningful impacts on public health. Novel strategies facilitated both by new tools (e.g., "next generation sequencing") and the revolutionary availability of digital sources of data related to food sales and consumption trends are being developed to contribute to the ability to (i) detect outbreaks and (ii) implicate the food vector causing the disease. However, methodologies that harness these new tools and data sources to contribute to solving part (iii) of the outbreak investigation, localizing the source, have not been brought to bear on the problem, despite the availability of a large amount of relevant system information, as we discuss below.

To fill this gap, this thesis develops novel methods that leverage currently unutilized or underutilized sources of information to contribute to the food safety regulator's ability to *identify the location source* of an outbreak. If public health officials can more quickly and successfully identify the location source of contamination, the outbreak can be stopped from spreading, the number of people who get ill can be reduced, and unmerited damages to the food industry can be avoided. This work therefore takes a deep look into the question, *how can we locate the source of contamination with greater accuracy, certainty, and speed?*

We see the opportunity to significantly decrease the time until source detection through working on two limitations: (1) deferring the investigation until confirmation of an outbreak has been established through PulseNet, and (2) failing to utilize all of the data available to contribute to solving the problem – and by association, developing methods to harness this data. Regarding (1), we acknowledge that this confirmation is a vital step to ultimately establishing with certainty the link between a case and an epidemic strain. However for the reasons discussed above, public health authorities will have access to these possible cases of illness as soon as a preliminary diagnosis is received at the site of medical care, which can be a few days to 3+ weeks before PulseNet confirmation. Furthermore, these initial, "tentative" cases could be supplemented by using case reports identified using Digital Disease Detection methods as mentioned above. We propose leveraging these initial case diagnoses to enable an earlier investigation of convergent sources of contamination.

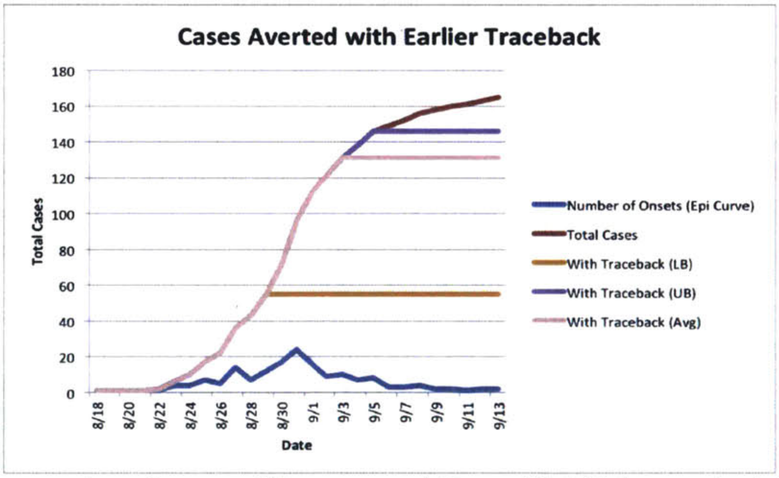


Figure 1.3. Representation of cases of illness potentially avoided if the traceback investigation had begun as soon as 10 cases had been identified as cases of E. coli O157:H7. The "Number of Onsets" and the "Total Cases" represent the actual outbreak data while the "With Traceback (LB, UB, Avg)" represent the lower bound, upper bound, and average number of cases avoided with a more timely recall. Source of data: (CDC 2006).

The value of an earlier start can be illustrated by examining the timeline of the foodborne disease investigation involving the 2006 outbreak of E. coli O157:H7 in fresh spinach, which caused 206 reported infections and 4 deaths across 26 states (CDC 2006, CDPH 2007). Based on the range of dates given in the timeline of case reporting (Figure 1.1 and Appendix 1.1) and the disease incubation period for E. coli O157:H7 (FDA 2016), Figure 1.3 represents the lower bound, upper bound, and average number of illnesses that may be have been avoided – 110, 19, and 34, respectively – if the traceback investigation had commenced by the time a cluster of 10 cases of E. coli had been preliminarily identified (***T_1*** in Figure 1.3) instead of waiting for confirmation of the cluster by PulseNet (***T_2*** in Figure 1.3). In reality, the regulatory traceback investigation did not begin until PulseNet confirmed the link between the outbreak and the cases, which turned out to be well after the epidemic curve had peaked (CDC 2006).

Regarding (2), we start off by noting that the triangulation approach to source identification mentioned above is fundamentally limited by its inability to leverage available information that can be contribute to identifying the location source of an outbreak. Because of the logistical limitations, investigators are only able to make use of a small subset of the reported cases of illness – data that serves as evidence in the source location problem. With only a few pieces of evidence, the time consuming traceback will often be unsuccessful in narrowing down the problem significantly. Furthermore, the process considers only supply chain structural information, that is, whether a link exists between different supply chain actors, while missing other dimensions of supply chain data that can help to differentiate between possible sources.

Moreover, triangulation represents a missed opportunity to utilize other valuable system information to solve the source localization problem. When each supply chain pathway is considered individually, the greater food distribution system that these pathways and their related supply chain actors are a part of is ignored.

1.3.3. A network approach to traceback

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 contamination distributed across a network can best be solved by considering this network structure and the dimensions of information it contains. The approach proposed and developed in this thesis is built upon the fundamental concept that we can utilize this network structure and its multiple dimensions of information better solve the problem of tracing the source of large-scale outbreaks.

While the exact parameters will vary from product to product, all food distribution systems can ultimately be represented by the same layered, directed network structure, as described in Chapter 2. The distribution network is made up of multiple stages of production, distribution, storage, and consumption, where each stage represents a specific class of supply chain actors. Food is created in the first stage, which represents the point of production at a farm or other type of producer, and is distributed along links by logistic service providers (represented by links) to interior stage nodes until it reaches the final stage, representing point of sale at retail or food service establishment. The interior stages may be involved with storage, collection, or further processing of the commodity. The source of a large-scale, multi-state outbreak will originate at a high stage of the network, e.g. in the producing or processing stages; only nodes in these stages are able to reach downstream nodes in geographically distributed locations. Case reports of illness are associated with the retail node at which the offending product was purchased.

With this structure fully 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. The feasible sources can thus be identified as the set of nodes in the producing or processing stage that share at least one network path to all contaminated retailer nodes. The set of feasible sources resulting from utilizing all case report evidence will be smaller than that resulting from the subset considered in triangulation.

Network structural information provides a first cut into the source identification problem, enabling us to identify the feasible sources of contamination. To differentiate between the feasible sources, we can leverage further dimensions of information available *within* the network. First, each link contains information about the quantity or volume of goods traded between supply chain actors. Second, time dynamics along the links provide additional information. Each network path, or collection of directed links and nodes from origination to point of sale, contains information about the distribution of time that a contaminated product could have taken to travel these steps. Combining information from the volume along links and the resulting temporal distribution along paths provides insight into who is more likely to have transmitted to whom. This insight can help us to discriminate among the feasible sources.

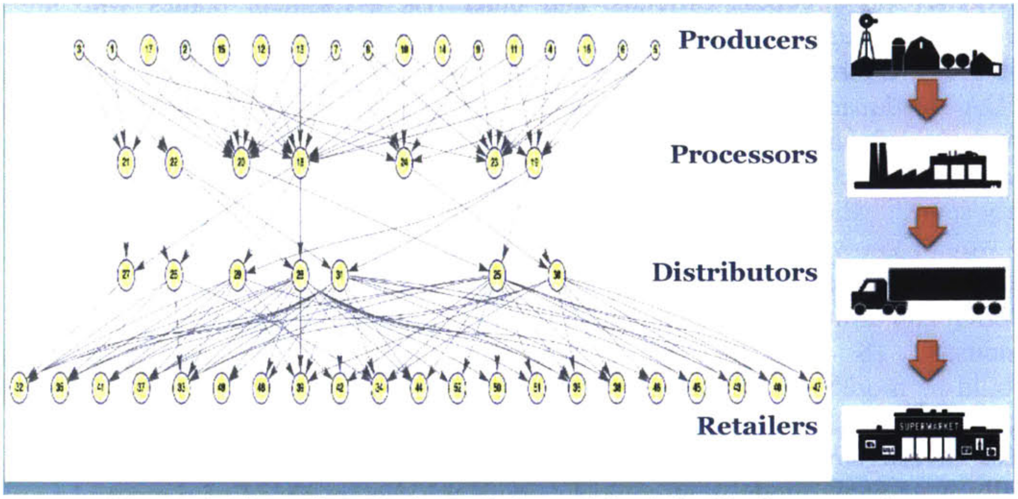


Figure 1. Illustration of a layered, directed food distribution network.

Our proposal

In this thesis, we develop a novel, network-theoretic approach to traceback that leverages all dimensions of information available for solving this problem. This approach can be used to supplement existing methods in outbreak investigation. We propose utilizing initial case diagnoses to enable an earlier investigation of convergent sources of contamination. In advance of sending out investigators and the onerous process of collecting official records to document product pathways, a low-cost, low-commitment computer model could be used to guide investigators specifically to the highest probability source locations while eliminating other locations as infeasible. This supplemental investigation would help to prioritize leads early on, potentially leading to faster resolution of the investigation.

A network approach to traceback presents important advantages in comparison with the current regulatory approach to traceback:

- **Immediate identification of *feasible* source locations**
- **Ranking of *feasible* sources by likelihood each one is *true source***
- **Straightforward use of ordered ranking to develop systematic investigation strategies**
- **Low financial and opportunity cost to implement**
    - Low financial cost since the traceback system would function on a computer model at very low cost to implement and no cost to operate.
    - Low opportunity cost because generation of investigation recommendations can be done at no exclusion of other approaches
- **Leverages all data available:**
    - All case reports, including initial or "tentative" cases; Comprehensive system network data

Due to these advantages, the proposed approach can offer significant improvements to the outbreak response process, such as the ability to

- **Outperform current triangulation methods**
- **Successfully resolve many more investigations**
- **Resolve investigations early enough that cases can be averted**

This thesis demonstrates these benefits, quantifying the specific improvement observed in computational results.

The communities we believe will benefit from the Traceback Tool are food-safety and public-health agencies, emergency preparedness officials, and other risk assessment bodies who are in need of improved methods for rapidly identifying the source of foodborne diseases. Sherri McGarry, the Foodborne Outbreak Coordinator in the Office of Compliance, Center for Food Safety and Applied Nutrition at FDA Headquarters and Erica Pomeroy, Consumer Safety Officer at the FDA San Francisco District Office, have emphasized the need for scientifically sound approaches to guide investigation and control measures. Mrs. McGarry has asserted that the most difficult time of the outbreak investigation is during the early stages "when there are many different potential suppliers who all fall into the same time frame, but no way of knowing who is implicated until commonalities start to appear...Any measure that will help to determine where we should focus our attention and give leads on the investigation would have a lot of application and utility for public health, and for business as the longer the outbreak the greater the impact on industry" (S. McGarry, personal communication, December 20, 2012). Furthermore, "messaging could be more targeted because we would be able to narrow down more quickly where the product is not coming from...This could really make a difference early on!" Ms. Pomeroy, a boots-on-the-ground outbreak investigator, believes our methodology would help her to direct an investigation: "If I could potentially have some model that could leverage risk factors ahead of time, and to combine this with whatever information exists about the nature of the distribution and exposure point, that would provide a great value" (Erica Pomeroy, personal communication, January 3, 2013).

Beyond its immediate and clear public health relevance, improving the ability to locate the outbreak source has real application to a current federal-level policy debate. In 2011, the landmark Food Safety Modernization Act (FSMA), the first major reform to food safety in 70 years, was signed into law. Among the Act's many provisions is an extension of traceability requirements. Specifically, FSMA extends the authority of the FDA to establish, as appropriate, "a product tracing system to *receive information that improves the capacity to effectively and rapidly track and trace food* that is in the United States or offered for import into the United States." The Act does not include specific provisions for implementing the law, and the regulatory timeline for determining the requirements is still developing. We suggest that the source localization framework developed in this thesis presents a viable approach to improving the capacity to "effectively and rapidly track and trace food."

1.4. Existing network approaches to traceback

We now consider how network information is leveraged by existing studies in the research literature to approach the problem of outbreak detection. Most work involving networks and outbreaks of contamination has been focused on the forward problem of understanding and forecasting the spreading process and its dependence on the structure of the underlying network. A portion of this research has been specific to the case of foodborne disease outbreaks on food distribution networks. In recent years, we have seen some work on the inverse problem of identifying the source of general outbreaks. Nonetheless, none of these studies have been specific to the case of foodborne disease. In the following, we review both forward contamination-spreading studies and inverse problems, focusing on the latter. We show that none of the existing approaches are designed specifically for tracing back the source of outbreaks on the unique structure of food distribution networks, highlighting how this makes them less effective in solving the problem.

1.4.1. Forward spreading problem

Recent years have seen the advent of epidemiological and network-theoretical approaches to the problem of identifying the geographical origin of large-scale outbreaks of disease. The first type of approach involves methods developed to determine the risk and characterize the impact of outbreaks. Computer and mathematical models are being used to simulate spreading dynamics in order to understand the potential impact of foodborne disease outbreaks on consumers. At the urging of the FDA commissioner in response to 9/11, Harlander and Sholl (2007) of BT Safety, LLC, created a simulation system to demonstrate the potential magnitude of public health and economic consequences of a specific contamination scenario, e.g., introduction of botulinum toxin into specialty ice cream production. More recent studies have used GIS (Geographical Information System) based spatial analysis of road networks to identify vulnerabilities and measure risks associated with contaminated food (Beni et al. 2011; Hashemi et al. 2012). Some modeling approaches have focused on the role of supply chain structure in determining spreading dynamics. Wein and Liu (2005) developed a mathematical model of a cows-to-consumers supply chain associated with a single milk processing facility that is the victim of a deliberate release of botulinum toxin. Their model evaluates the dispersion of contaminant and subsequent impact on consumers resulting from various types of flows and supply structures. Pinior et al. (2012) used simulation to explore the extent to which inter-dairy connections influence the spatial spread of contaminated milk and resulting contamination risk. Conrad et al. (2012) present a general methodology for the stochastic mapping of fresh produce distribution networks and illustrate an application to a small-scale supply chain case study. This study is the first to introduce the idea that mapping network structure can help to determine the sources of a contamination through backward tracing.

1.4.2. Source identification problem

The second type of approach involves methods for tracing back outbreaks to the source; these can be used to supplement or assist investigators in identifying and narrowing down potential sources of contamination. Attention to understanding and tracing back foodborne disease outbreaks has emerged in academic research communities only very recently, but there has been significant effort in studying the dynamics of outbreaks on networks more generally over the past couple decades (Moore and Newman 2000; Pastor-Santorras and Vespignani 2001; Newman 2002; Keeling and Eames 2005; Riley 2007; Lind et al. 2007; Brockmann et al. 2009). This research has focused on the forward problem of understanding and forecasting the diffusion process and its dependence on the structure of the underlying network. Very little work has been done on the backward-tracing problem of identifying the source of an epidemic. Shah and Zaman (2012) developed a maximum likelihood detection estimator for inference of the unknown source for general graphs, which assumes a fully specified transmission network with precise transmission times. Pinto at al. (2012) built upon this approach, assuming only a small fraction the nodes in the transmission network can be observed. Their approach relies on the precise transmission timing at the observed nodes to fill in the unknown components of the infection graph. Farajtabar et al. (2015) introduce a new approach that improves upon the detection ability of Pinto et al. by developing a model that represents a given outbreak scenario with more specificity. Their approach is based on a two stages, the first of which is to fit a model of contamination diffusion to a set of multiple historical outbreak datasets. Second, they identify the source as the node that maximizes the likelihood of the observation times and locations associated with an ongoing outbreak according to the learned diffusion model. Comin and da Fontoura Costa (2012) demonstrated the relationship between network centrality measures – standard degree, betweenness, closeness, and eigenvector centrality, and the source of spreading. Centrality is also key to the methods used by Prakash et al. (2014) and Fioriti and Chinnici (2012), who developed spectral techniques to identify a set (not a single source) of origin nodes.

Whether it be disease in human contact networks, rumor outbreaks in social media, or viruses in computer networks, these studies all assume knowledge that would not be available during actual epidemic outbreaks of foodborne disease. First, all approaches assume that contamination observations are available as scattered throughout the entire cascade process, with transmissions observed both close and far removed from the source (in network hops) and make use of this information to zero-in on the contamination start time and location. The information assumed is good enough either to form or to estimate the exact transmission network/tree of who passed the contamination on to whom. However in outbreaks of foodborne disease, the only temporal data available to bound the determination of when the contamination passed through certain network points is at the very end of the trajectory. This is because a contamination is only observed when a consumer reports an illness, whereas the contamination will traverse multiple network steps through supply chain nodes before finally making it to that downstream consumer. Temporal data available only at the very end of a contamination's trajectory does not provide enough information to reconstruct the transmission tree/network. Methods assuming distributed transmission time data thus cannot be applied to the case of foodborne disease. Additionally, even if the temporal data on case reports is available as scattered throughout the contamination cascade as is the case of outbreaks of infectious disease spreading from human to human, the assumption that the temporal information available will be good enough to assemble the transmission tree is rarely the case in practice. Real outbreaks are characterized by significant underreporting, with CDC estimates of underreporting varying from 10 to 75 times for different pathogens (Scallan et al. 2011). Application of these methods to traceback the source of these outbreaks in real-time, given knowledge about the diffusion network available at the time of the emerging event, have yet to be seen.

However while the temporal data may be reduced in the case of foodborne disease, other information is available: the relative amounts of commodity volume traded across the links of the *weighted* distribution network. A single approach to apply network-theoretic methods to identifying the location origin of an outbreak of foodborne disease, which accounts for the unique elements of food distribution networks, is presented in Brockmann and Helbing (2013) and Manitz et al. (2014). These studies apply a network-theoretical method for source localization in general complex networks to the context of foodborne disease, demonstrating the ability to resolve the source of the application to the 2011 E.coli outbreak in Germany (WHO 2011). A spatial transport network representing travel paths between geographical districts underlies their approach, which is focused on identifying the geo-spatial location of outbreak origin rather than its location within the supply chain. Their method, called the "effective distance," leverages the observation that while a contaminant can travel a multitude of paths to any other node, the dynamics of transmission are dominated by (i) the shortest paths and (ii) the highest probability paths; correspondingly, longer, lower probability paths are penalized (Brockmann and Helbing 2013). Its application requires only information on case reports collected by public health institutions rather than a detailed history of transmission. Applied to the historical case, the effective distance method is able to localize the outbreak source within a narrow geographical area, dramatically reducing the set of possible origin locations. While the traceback resolution is promising, the method focuses only on the network-spatial predictive component and does not leverage predictive information to be gained from considering the times of illness reports or contribute to the problem of identifying the start time of the epidemic. In addition, it is ultimately limited by the specificity and resolution of the underlying network, which models general transport systems and does not differentiate between supply chain actors, allowing any node in the network to be a possible source of spreading. The features of the multi-scale structure of transport networks for which the method is designed – great heterogeneity in degree distribution and length of paths travelled by contaminations – are not typical of many supply chain networks; multi-partite networks for which the length of all network paths from source node to contamination point are equal or close to the same number of steps, differing at most by the number of layers in the network.

To summarize, multiple practical constraints distinguish this problem from source identification in other network contexts: (i) that only a small fraction of illnesses are reported, (ii) that the reported times are imprecise, and (iii) that the presence of contamination at locations within the distribution network is unknown or hidden; thus, the source of contamination can be recovered only from the information associated with the reported illnesses. Any methodological approach to this problem should be developed around these constraints. Furthermore, multiple dimensions of information are available to the foodborne disease outbreak detection problem, and these should be leveraged: directed, layered network structure;

the temporal dynamics of how contamination spreads; and weighting information provided by volumes of commodity traded. While many successes have been presented in the literature, an approach has not been developed that is (i) specific to the context of foodborne disease and (ii) leverages and combines the multiple sources of data available to this problem. The network structure assumed in all existing work is a general, undirected network in which any node can be the source and any node can be contaminated, which is not the case for food distribution networks. Furthermore, the approaches can be categorized as utilizing either network structure and volume information or network structure and temporal information, but not a combination of the two.

1.4.3. Network Structure and Traceback Accuracy

Network structure is a key determinant in spreading dynamics (Newman 2002; Pastor-Satorras and Vespignani 2001; Gonzalez et al. 2009; Onnela et al. 2011; Grady et al. 2011), and it should likewise be a determining factor in backward tracing. One can easily imagine that for a network composed of vertically integrated supply chains, any observation of contamination can be correctly traced back to the original source. On the other hand, if there are a lot of cross distribution links among entities in the chain, the uncertainties of source contamination can be extremely complex, e.g., if tracing is investigated on the Sara Lee Bread supply network discussed above. The more cases showing contamination, the fewer locations are suspect due to the topological properties of the network. That is, only so many farms could have caused contamination at precisely this set of contaminated retailer locations. Understanding the role of network structural parameters in affecting the ability to identify the source has important practical implications, helping to develop an understanding of how and in for what situations we expect to achieve accurate results. While this is an important area of research, our review has not identified any studies that systematically explore the role of network structural properties in determining the ability to identify the source of spreading phenomena for complex networks in general, let alone food distribution networks.

1.5. Research Objectives

This thesis aims to develop an implementable framework and set of methods for real-time source detection that is specific to the context of foodborne disease, filling the literature gap identified above. Specifically, this research seeks to address the two following objectives:

(1) **Develop a framework and approximate inference approach for solving the constrained foodborne disease source identification problem.** We design a solution that is specific to the problem context. We assume an accurate representation of the supply chain network structure as a directed, layered network from farm through processing, packaging, and distribution, to retail, where only certain node types can be considered as the source (e.g. farms or processors), and only consumer nodes can report infection. In addition to network structural information, the approach leverages the spatial and temporal dimensions of problem while explicitly incorporating the uncertainty intrinsic to the evidence available to investigators at the time of an outbreak.

(2) **Explore how the structural properties of food distribution networks affect the ability to accurately identify the source of contamination**, ultimately enabling us to determine for what combination of structural features and thus what food types our methodology is likely to provide the greatest benefit. We identify specific features of food distribution networks that play a greater role in determining the traceability of an outbreak. We build a modeling framework that incorporates these features, coming at the problem from two ends. We develop models representing stylized versions of the problem; these will help us to derive new, general insights regarding how structural properties of food distribution networks affect the ability to infer the source of a contamination. From the other direction, we develop geographically and structurally realistic regional models that include true system complexity, which will allow us to validate and

test robustness of stylized results, to quantify the potential benefits of the investigation strategies being developed. This analysis will help us to determine the data requirements of the framework, considering the level of detail in in modeling and data necessary to achieve high traceback performance without oversupplying it.

(3) **Develop investigation interventions to guide investigators at the tactical level in applying their search effort to solve an investigation and minimize impact on public health.** Mechanisms are developed to support investigators in deciding (i) when and where to **deploy investigators** and (ii) when to **message the public** implicating the likely outbreak source(s) and what locations it should include. The procedures involve investigators predetermining a desired accuracy level and allocating a non-monetary "budget" of resources to the investigation.

1.6. Thesis Contributions and Organization

1.6.1. Holistic source detection system

The practical objective of this thesis is to contribute to the food safety regulator's ability to efficiently locate the source of large-scale outbreaks while contamination-caused illnesses are occurring, thereby resolving investigations earlier and averting potential illnesses. To this effect, the primary contribution of this thesis is the development of a holistic system for rapid identification of the source within the constraints of available or acquirable resources. In advance of sending out investigators and the onerous process of collecting official records to document product pathways, this system would be applied to effectively prioritize investigation leads. It would be used to definitively identify *feasible* source locations, and would narrow the *true* source to within a small, bounded set of high probability candidates with a quantified accuracy. This bounded set of candidates would inform specific recommendations regarding when and to which facilities to implicate in public service messaging and to deploy investigators to confirm predictions. It would be accomplishable at marginal cost, since the traceback system would function on a computer model at low cost to implement and no cost to operate.

On a high-level, the system is based on a network-theoretic, spatio-temporal framework for outbreak detection that functions on current domestic food systems network data, reported cases of illness and a prior probability function over likely sources. The traceback framework outputs a set of feasible sources of contamination and their relative probability of being the true source, producing an ordered ranking over the feasible candidates. This source data is input into a decision model for investigation interventions, which provides investigators with a set of recommendations for two distinct options for action: (i) where to send on-the-ground investigators and (ii) when and what to message to the general public about potential contaminated foods. The development and review of this holistic framework is provided in the four main chapters of this thesis and a future plan for practical validation is provided in the conclusion.

Figure 1.4 provides a stylized flow diagram of the holistic framework, broken into 8 stages. Stages 1 – 6 follow a linear flow, starting with the necessary externally generated data inputs and finishing with dual-decision making recommendations for food-safety regulators. Stages 7 and 8 are a data generation and feedback cycle that updates the accuracy of traceback and recommendation. Stages 1 – 3 are addressed in Chapter 2; the accuracy of Stage 4 is addressed in Chapter 3; Stages 5 – 8 are the focus of Chapter 4; how to implement the entire framework is addressed in Chapter 5.

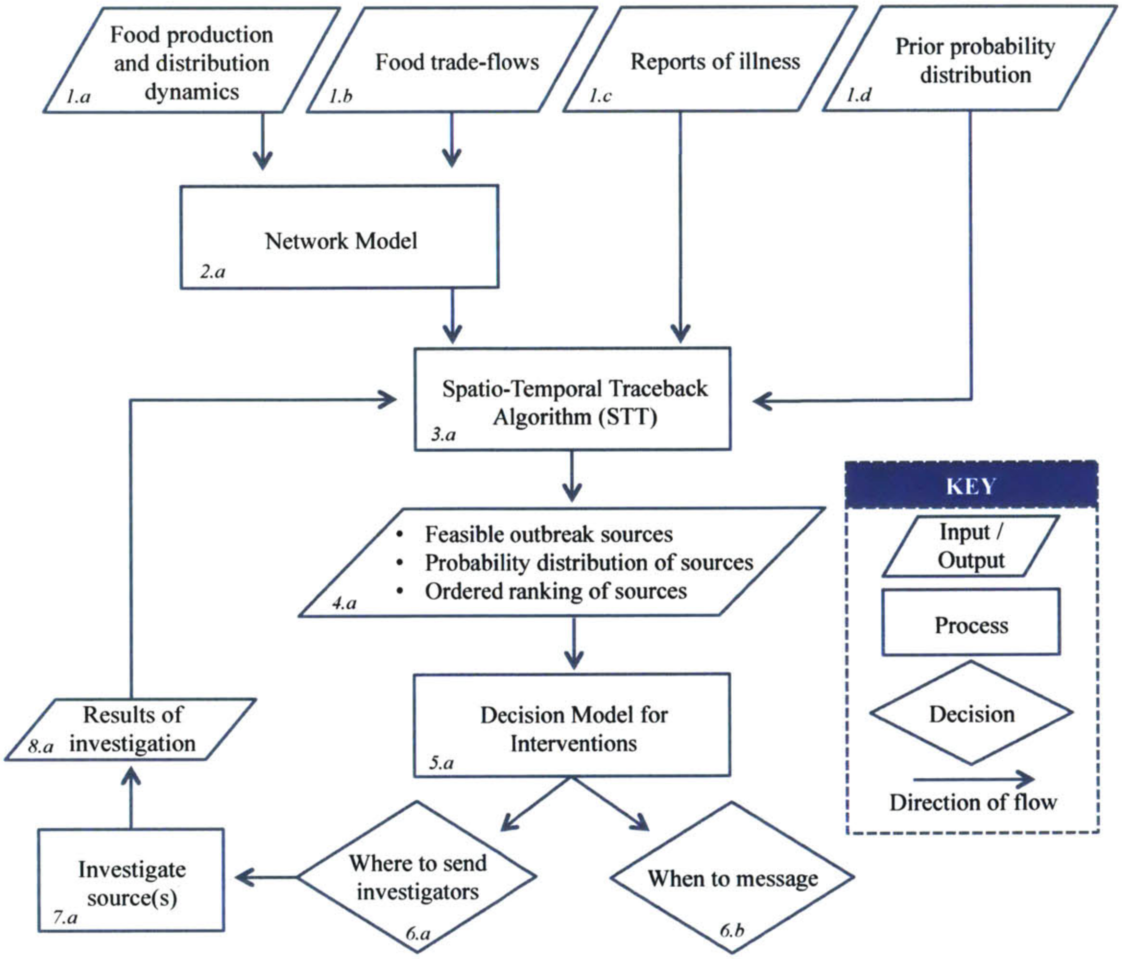

Food production and distribution dynamics
1.a
Food trade-flows
1.b
Reports of illness
1.c
Prior probability distribution
1.d
Network Model
2.a
Spatio-Temporal Traceback Algorithm (STT)
3.a
• Feasible outbreak sources
• Probability distribution of sources
• Ordered ranking of sources
4.a
Decision Model for Interventions
5.a
Where to send investigators
6.a
When to message
6.b
Investigate source(s)
7.a
Results of investigation
8.a
KEY
Input / Output
Process
Decision
Direction of flow

Figure 1.4. Flow diagram of holistic source traceback system.

1.6.2. Challenges and Limitations

Vital to using the traceback framework during actual contamination events is a real-world supply chain network model for any food involved in a contamination event. Furthermore, the ultimate accuracy of the combined model-and-method approach will depend on the fidelity of the network data. The network model must take a systemic view for a given commodity, since the national food supply is a coherent system; products, regions, and actors are connected and are often interdependent. To implement the traceback methodology in an emergency, a database of network models for various food types would need to be constructed in advance so that the methodology is ready to launch immediately. However as addressed in Section 1.2, data on structures of aggregated distribution networks is not readily available from publically held sources, nor does it exist in any database in an organized form. Acquiring and organizing this information at this level of detail presents three major challenges. First, food distribution networks are markets characterized by inherent stochasticity. Some trade relationships are enduring while others may be based on transitory spot-markets. Second is the challenge of acquiring and organizing the information. Trade relationships are proprietary information that businesses are often reluctant to share, since they constitute competitive information that presents advantages in the low-margin retail food industry. While each facility must participate with at least the "one-up, one-back" recordkeeping mandated by the Bioterrorism Act of 2002, this data is proprietary and must be shared with FDA only when responding to an incident. Third is that even if this information is made available to regulators, collecting and organizing the data in a well-maintained system would present extensive practical data-management challenges. Foods of today are complex and outbreaks can occur in foods containing dozens of ingredients.

In Chapter 5 we consider these challenges in detail and ultimately posit an implementable traceback system that would have limitations in its granularity. For the purposes of developing our methods we assume in Chapters 2 – 4 that we have access to well-characterized food distribution data.

1.6.3. Thesis organization

Chapter 2 is the main theoretical contribution of this thesis. We present a network-theoretical framework for source localization during foodborne contamination events. As discussed in the previous section, multiple practical constraints distinguish this problem from source identification in other network contexts. We develop an approach that is specific to the problem of foodborne disease, accounting for the complexities distinguishing it from source identification in other network contexts.

We introduce an approximate inference algorithm for solving the constrained foodborne disease source identification problem, called *Spatio-Temporal Traceback* (STT). Given the distribution and timing of outbreak-related cases and knowledge of the distribution network, characterized by volume flow proportions between nodes at known geographic locations, the inference algorithm uses probabilistic induction and network analysis to determine the probability that any location is the outbreak source. It returns a probability mass function (PMF) representing the posterior probability that any feasible candidate is the true outbreak source, and an estimate of the outbreak initiation time. Due to topological properties of the network, the greater the number and dispersion of cases, the fewer locations are suspect; the greater the homogeneity of the network structure, the more differentiated the resulting probabilities will be. The feasible set is rank-ordered according to their probability values, identifying the estimate for the likeliest source or set of sources. Underlying the method is an assumed structural model of the distribution network and a diffusion model of contamination. Given the network and diffusion model, the method utilizes information from both the location and time of each illness report to determine (i) its most likely trajectory through the network and (ii) the aggregate probability across all possible trajectories, for each feasible source node. It then proposes a "two for the price of one" solution, which simultaneously identifies the most likely contamination source and its initiation time by maximizing the likelihood of the observed contamination times and aggregate path probabilities.

We develop a simulation-based evaluation framework and set of accuracy metrics that enable us to measure the success of the traceback methodology across a wide range of network structures and outbreak scenarios. We then present an initial performance evaluation study, applying the techniques illustrative numerical examples in order to provide (i) an indication of the accuracy and applicability of the method and (ii) a first step towards understanding how the accuracy of detection depends on the structural property of heterogeneity of a network and on the stochastic evolution of the disease trajectory. These analyses provide insights derived from this initial evaluation and the implications for practice.

The contributions of Chapter 2 fall into four categories:

i. We formulate an original network-theoretical framework for identifying the location and initiation time of the source of a large-scale outbreak of foodborne disease, which is specific to the unique complexities distinguishing this problem from source identification in other network contexts.
ii. We develop an approximate solution algorithm that exploits the temporal and structural dimensions of this problem.
iii. Through illustrative numerical examples, we provide an indication of the accuracy and applicability of the method and show that it follows intuitive properties.

iv. We demonstrate that the algorithm's convergence process depends only on the network topology and is independent of initial conditions, i.e., does not depend to the choice of prior distribution over the sources.

Our ultimate objective is to evaluate the utility of these methods in real-world scenarios, and in comparison to existing methods in outbreak investigation. **Chapter 3** provides a first step towards validating the accuracy and applicability of the traceback methodology in real-world scenarios. First, a generalizable modeling framework representing key structural and spatial features characterizing real distribution networks is developed and implemented to model the distribution of two specific commodities in the US: tomatoes and lettuce. We evaluate the performance of the traceback methodology applied to these two realistic structures, demonstrating high accuracy in both identifying and localizing the outbreak source. We analyze results to provide an understanding of how the successful resolution of an outbreak depends on the structure of a network, and the robustness of results for food distribution networks in general. We then develop a framework for quantifying benefits in comparison to existing approaches to foodborne disease outbreak source detection: to a practical baseline meant to demonstrate current methods applied in practice during outbreak investigations, and furthermore, and to a best-in-class method presented in the literature (Brockmann and Helbing 2013). Our method significantly outperforms both heuristics and the state of the art theoretical method across a wide range of outbreak structures. Theoretically, these results demonstrate the suitability of our specific methodological approach to the problem of localizing the source of foodborne disease outbreaks. Practically, they suggest that our method can contribute to the traceback investigation process; a conclusion that will ultimately need to be validated beyond the simulation results presented here.

The contributions of Chapter 3 fall into four categories:

i. We identify key structural and spatial features characterizing real distribution networks and develop a generalizable modeling framework representing these features utilizing available data.
ii. We show the traceback methodology robustly identifies and localizes the outbreak source when applied to realistic network structures; these results serve as a first step in validating its application in practice.
iii. We derive generalizable conclusions regarding the dependence of traceback accuracy on network structural features.
iv. We show that the method can result in significant benefits in accuracy and efficiency when compared with existing approaches in foodborne disease outbreak source detection: a practical heuristic meant to model current methods applied in practice, and a best-in-class method presented in the literature. These results demonstrate the suitability of our solution to the problem of localizing the source of foodborne disease outbreaks.

The algorithm output can form the basis of a decision-making tool enabling public health and emergency preparedness officials to facilitate a probabilistic analysis of the source of an ongoing outbreak, and to more effectively allocate investigative and communication resources. In **Chapter 4**, we develop strategies for real-time investigation response based on the traceback inference methodology, which include determining (i) when to message the public about the outbreak source and how to frame the statement, and (ii) when and to which potential source candidates to send out investigators to confirm predictions. We quantify potential benefits to public health, industry, and investigators by implementing the traceback methodology together with this decision model. These benefits are quantified according to the improvement in accuracy, certainty, and speed of the traceback analysis and the resulting reduction in the number of illnesses, cost of investigation, and the impact on the industry.

The contributions of Chapter 4 fall under four categories:

i. We define the attributes accuracy, benefit to public health, and cost to regulators or industry as characterizing the performance of investigation interventions; these performance measures allow us to define an framework for enumerating, quantifying, and comparing intervention options.
ii. We propose mechanisms based on the traceback methodology of this thesis for deciding when and where to deploy investigators and when and what to message to the public given an allowable level of risk and the resources available.
iii. We quantify the potential benefits to public health possible if interventions based on these mechanisms are implemented, measured in terms of illnesses averted.
iv. We show from computational results that these methods demonstrate great potential to improve upon current methods in outbreak response, recommending whether, when, and with what to respond during an outbreak.

To implement the traceback methodology in an emergency, a database of network models for various food types would need to be constructed in advance so that the methodology is ready to launch immediately. The feasibility of implementing the traceback methodology and the performance of the combined model-and-method approach will ultimately depend on the properties of the underlying network model and the data informing them. In **Chapter 5** we consider multiple approaches for modeling the supply chain network and implementing the bank of network maps. We examine the potential accuracy of each alternative, considering in particular the level of detail necessary to achieve high traceback performance without oversupplying it. We suggest a means to collect the necessary data and discuss the feasibility of its implementation. On the basis of these analyses, we recommend a combined model-and-method approach that would form an implementable system for real-time source detection and suggest next steps in its evaluation and implementation.

The contributions of Chapter 5 fall in four categories:
i. We specify the requirements of the underlying system-wide supply chain network model and propose four approaches for modeling the structure meeting the necessary requirements.
ii. We examine the potential accuracy of each alternative, considering in particular the level of detail necessary to achieve high traceback performance without oversupplying it.
iii. We suggest a means to collect the necessary data and discuss the feasibility of its implementation.
iv. On the basis of these analyses, we recommend a combined model-and-method approach that would form a ready-to-implement system for real-time source detection and suggest next steps in its evaluation and implementation.

It is important to stress that the results derived and benefits quantified in this thesis are estimated from simulation; live use of these techniques has yet to occur and may demonstrate features of the real problem inadvertently omitted from the modeling. Extensive testing of the methodology will be necessary to determine the utility to public health, measured in terms of how much earlier an investigation can be resolved and how many illnesses averted as a result. In **Chapter 6**, we conclude the thesis by summarizing the results of research and recommending a specific, rigorous, two-stage evaluation the traceback methodology. The first step will be to demonstrate the ability to correctly localize the origin of historical outbreaks. A research project undertaken this activity has been planned for October – December 2016, described in detail in the Conclusion. This is one step towards validation, but still within the confines of research. The data available post-fact will always be better than what would be available at the time of an outbreak, due to the delays and inaccuracies in case reporting. Real-time application of the tool during outbreak emergencies will ultimately be necessary. Upon the success of the theoretical application to historical cases, follow-on studies should be conducted to expand the validation process in live experiment. We are hopeful this approach can significantly improve upon current methods and thus increase the capability of investigators to quickly and efficiently respond to food safety problems.

Appendix 1.1
Timeline for Case Reporting

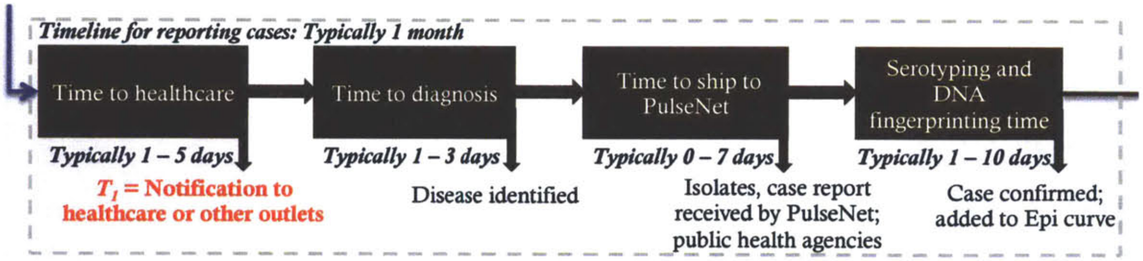

Timeline for reporting cases: Typically 1 month
Time to healthcare
Time to diagnosis
Time to ship to PulseNet
Serotyping and DNA fingerprinting time
Typically 1 – 5 days
Typically 1 – 3 days
Typically 0 – 7 days
Typically 1 – 10 days
T_1 = Notification to healthcare or other outlets
Disease identified
Isolates, case report received by PulseNet; public health agencies
Case confirmed; added to Epi curve

Figure A1.1. Timeline for Reporting Cases. Data sources: (CDC 2015a-b, FDA 2001, Wilkins et al. 2015, WHO 2008).

After a patient begins to experience symptoms, a series of events must occur before public health officials can confirm that the patient is part of an outbreak. These events, known together as the Timeline for Reporting Cases, mean there can be a substantial delay between the start of illness and confirmation that a patient is part of an outbreak. The timeline is as follows (CDC 2012):

- **The time to healthcare** is time from the first symptom until the person seeks medical care, when a stool sample is collected for laboratory testing. This time will depend on the intensity and the duration of symptoms and is typically 1-5 days, but can be longer.
- **The time to diagnosis** is the time from when a person provides a sample to when the result is obtained from a laboratory, either off-site on-site at the location of medical care. This may be 1-3 days from the time the sample is received in the laboratory.
- **The time to ship to PulseNet** is the time required to ship the isolated pathogen from the diagnostic laboratory to PulseNet, the national network of public health and food regulatory agency laboratories coordinated by the CDC that perform serotyping and standardized molecular subtyping ("DNA fingerprinting") of foodborne disease-causing bacteria to distinguish strains at the DNA level. This can take between 0-7 days depending on transportation arrangements within a state and the distance between the clinical laboratory and the closest PulseNet laboratory. Diagnostic laboratories are not required by law to forward isolates to PulseNet labs, however.
- **The time to serotyping and "DNA fingerprinting"** is the time required for PulseNet to serotype and to perform "DNA fingerprinting" on the isolate and compare it with the outbreak pattern. Serotyping may take up to 3 days and "DNA fingerprinting" can be accomplished in 24 hours, however in practice this process may take 1-10 days. If a match is found, then it is at this stage that the patient is confirmed as part of an outbreak and the case is added to the count of cases, known as the epidemiological curve or epi curve. Each case is identified in the epi curve according to the patient's recalled date of symptom onset.

As a result, the delay due to the timeline for reporting cases can range from three days to three weeks. Case counts in the midst of an outbreak investigation are therefore always preliminary and must be interpreted within this context.

Chapter 2: Traceback Methodology

In this Chapter, we develop a framework and approximate inference approach for solving the source identification problem, given knowledge of the underlying distribution network and the set of observed illnesses at specific network locations and reported times. As discussed in Chapter 1, multiple complexities distinguish this problem from source identification in other network contexts: (i) that only a small fraction of illnesses are reported, (ii) that the reported times are imprecise, and (iii) that the presence of contamination at locations within the distribution network is unknown or hidden; thus, the source of contamination can be recovered only from the information associated with the reported illnesses. Any approach to this problem should be developed around these complexities. Another feature distinguishing this problem and that can be exploited in its solution is the observation, discussed in more detail in Chapter 3, is that real food distribution network structures often exhibit considerable variability in the distribution of links and the flow proportions. A solution designed for this condition, without loss of generalizability to cases when differentiation is less pronounced, will perform better across a greater variety of scenarios that might present in practice.

There are multiple ways to approach the constrained foodborne disease traceback problem. Our work developed and investigated two novel methods: *Bayesian Network Traceback* (BNT) and *Spatio-Temporal Traceback* (STT). In BNT, a time-independent approach is taken to identify the source based on the structure of the distribution network and the reported locations of contamination. The key idea is to view the problem from the perspective of a probabilistic graphical model, with a random variable that represents, for each node, whether the contamination has passed through that site (of food production, distribution, or retail). The graphical model characterizes, through a set of conditional probability distributions, how the observation of a contamination at a given node increases the probability that the contamination has traveled through adjacent upstream and downstream nodes. Each random variable is binary, meaning a node can represent one of two statuses: contaminated or not contaminated; thus, BNT leverages "negative" information from nodes known not to present the contamination, but does not incorporate extra information from multiple illnesses at the same node. A formulation is constructed that uses the graphical model to identify the source of contamination as the node that maximizes the joint likelihood of the set of reported nodes. Spatio-Temporal Traceback utilizes information from both the location and time of each illness report individually, assuming the report times are accurate within a given uncertainty. An underlying contamination diffusion model is assumed, and the most likely source and initiation time are identified by maximizing the likelihood of the time delay densities of the observations given the diffusion model.

After implementing both methods and evaluating their performance in a variety of sensitivity tests, we find that unambiguously, STT performs better in two important dimensions: accuracy and computational scalability. Despite the theoretically accurate probability formulation modeled in BNT, STT achieves better accuracy by leveraging additional information. First, it incorporates temporal information from the observations, which despite the imprecision in their measurements, should not be disregarded. Second, it considers multiple cases of illness at each retailer, weighting pathways between source and observation node in proportion to the number of occurrences. Given its clear superiority, this thesis will focus on STT and its ability to localize the source to a well-defined region or a single node. A theoretical explanation of BNT can be found in Appendix 2.1.

The contributions of this chapter fall into four categories:

- We formulate an original network-theoretical framework for identifying the location and initiation time of the source of a large-scale outbreak of foodborne disease, which is specific to the unique complexities distinguishing this problem from source identification in other network contexts.
- We develop an approximate solution algorithm that exploits the temporal and structural dimensions of this problem.
- Through illustrative numerical examples, we provide an indication of the accuracy and applicability of the method and show that it follows intuitive properties.
- We demonstrate that the algorithm's convergence process depends strongly on the network structure and much less strongly on initial conditions, i.e., on the choice of prior distribution over the sources.

The chapter is organized as follows. In Section 2.1, we formulate the spatio-temporal framework and approximate algorithm for solving the source identification problem. In Section 2.2, a probabilistic simulation approach involving generalized food distribution network models and diffusion models of contamination is developed to evaluate the performance of the methodology. In Section 2.3, we subject the technique to an initial performance evaluation study. Section 2.4 concludes.

2.1. Spatio-Temporal Traceback Framework

2.1.1. Method Overview

In this section, we introduce the *Spatio-Temporal Traceback* framework and inference algorithm for solving the foodborne disease source identification problem. The framework requires the following input data:

- Food supply chain network information
    - o *Identity* of supply chain nodes and *location* in geography
    - o The *existence* of trade links between supply chain nodes and *volume* traded
    - o *Time dynamics* of how contamination spreads across the network
- Case report data:
    - o *Location* in the network
    - o *Time* of occurrence, according to patient's recalled time of illness onset

This input data provides multiple dimensions of information that can be used to solve the source identification problem: the network structure, volume, and temporal dynamics.

The algorithm first performs a "preprocessing" step, using the network structure to determine the feasible sources as the set of processing or producing stage nodes that share at least one network path to all contaminated nodes. For any given instantiation of the algorithm, the source must be assumed to be in either the processing or producing stage.

With the feasible set identified, the algorithm then determines the probability that each feasible source is the true source, given the observations of illness at specific node locations and times. To determine this probability, this chapter provides a derivation that factors out the volume-based probability contribution, decomposing the probability of being the true source into a volume component and a temporal component. The result is that the *posterior* probability of being the true source is the product of the Bayesian prior probability, a volume-based probability factor, and a temporal probability factor (Figure 2). An approach to efficiently estimate the volume and temporal probability factors, accounting for the computational constraints of operations on networks, is designed in the following sections.

The methodology determines the probability factors for prior, volume, and time, applying the same process for each feasible source node individually. The prior probability is informed by information external to the network structure. If information on known risk factors or expert opinion is not available, the relative production quantity at each feasible node is used, assuming that any product produced is equally likely to generate contamination *a priori*. The volume contribution quantifies the probability that a feasible source node could have reached all contaminated nodes. Here, we are essentially assuming that the (relative) total volume of goods flowing from the source to all contaminated nodes is as a proxy for this probability. We calculate this term using the weighted adjacency matrix representing the network. The temporal contribution quantifies the probability that the feasible source generated the observed illness times, given what we know about the time dynamics from contamination origination to observation. In other words, this is the probability that the feasible source node can "explain" all of the observed contamination times. To determine this probability we first identify the set of highest probability paths from the feasible source to each contaminated node. We then find the start time that maximizes the likelihood of the observation times – and record the associated probability, which is the area under the likelihood curve within a small "uncertainty window" around the observed contamination times.

Once each probability factor has been determined for a feasible source node, we multiply these probabilities together to determine the posterior probability that node is the true source. We do this for each feasible source node, normalizing the resulting set to define a posterior probability mass function (PMF) representing the probability-ordered ranking over the set of feasible sources. An illustrative PMF and ordered ranking is depicted in Figure 3. It also returns an estimate of the outbreak initiation time. Due to topological properties of the network, the greater the number and dispersion of cases, the fewer locations are suspect; the greater the variance in the network structural parameters (e.g. the number of other nodes each node is connected to; the distribution of volume along links) the more differentiated the resulting probabilities will be.

Underlying the method is an assumed structural model of the distribution network and a diffusion model of contamination. In the following, we develop the underlying network and outbreak contamination models, and then present the STT approach to source detection.

2.1.2. Network Model

Below, we develop the food distribution network model that serves as the foundation for the traceback methodology. A more thorough description of the key features and structural components of food distribution networks is provided in Chapter 3.

Food distribution networks

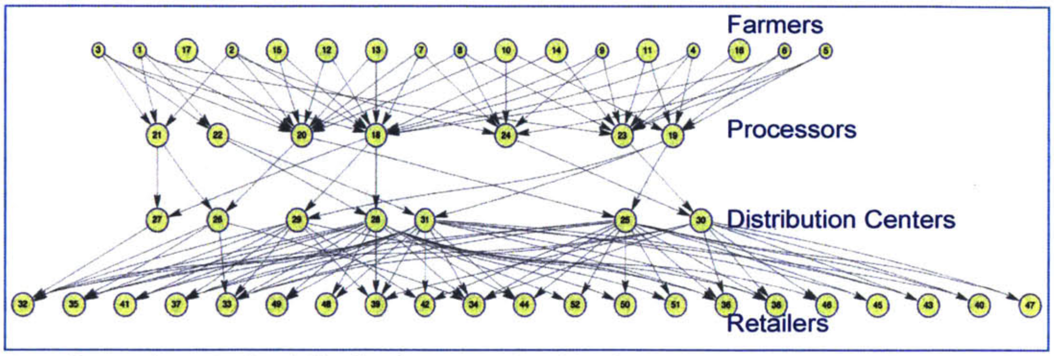


Figure 2.1. Illustration of a food distribution network with of 4 stages, or groups of supply chain actors, categorized into Farmers, Processors, Distribution Centers, and Retailers. This network consists of 17

Farmers, 7 Processors, 7 Distribution Centers, and 21 Retailers. Food is created by the Farmers, turned into products by the Processors, stored at Distribution Centers, and sold at Retailer nodes. Products are distributed from Processors to Distribution Centers and Distribution Centers to Retailers by logistic service providers (represented by links).

A food distribution network represents the aggregated (i.e. multi-company) distribution chain for a given commodity. A visual example of the structure of a food distribution network is provided in Figure 2.1. The network is made up of multiple stages of production, distribution, and storage or consumption, where each *stage* represents a specific class of supply chain actors. Food is created in the first stage, which represents the point of production at a Farm or other type of Producer, and is distributed along links by logistic service providers (represented by links) to interior stage nodes until it reaches the final stage, representing point of sale at Retail or Food Service. The interior stages can be involved with storage, collection, or further processing of the commodity. The network in Figure 1 is composed of 17 Farms, 7 Processors, 7 Distribution Centers or Warehouses, and 21 Retailers. We note that while supply chain actors within a stage generally trade only with actors in another stage, flows of product within stages and across multiple stages can occur. We do not explicitly account for these cases in our modeling analysis, but we note that the methodology developed here can represent these structures without any modifications. (Hashemi Beni et al. 2012, LeBlanc et al. 2015, Pinior et al. 2012, 2014).

Network model

We model the food distribution network as a directed, acyclic, N-partite graph $G\{V,E,N\}$. G is assumed to be known, though information on the status of edges may be missing. Let $V(n)$ denote the set of nodes u in stage $n = 1,...,N$. Directed edges of the form (u,v) in E may exist only between some $u \in V(n)$ and some $v \in V(n+1)$. The volume of food created per unit time at nodes $u \in V(1)$, normalized across all u, is denoted by f_u. F_0 is then a matrix of dimension $|s| \times |s|$ with diagonal elements equal to the normalized initial volumes f_u. The flow f_{uv} quantifies the average proportion of food sent from u to v per unit time, normalized across all outgoing links from u, along distance d_{uv}. F is a transition probability matrix of dimension $|V| \times |V|$ composed of the normalized flow proportions f_{uv}, such that $[F]_{u,v} = f_{uv}$. D is a matrix of the same dimension composed of the distances d_{uv}. Elements along the diagonal of F correspond to self loops and are thus equal to 0, with the exception of the flows corresponding to the retailer nodes $w \in V(N)$, which are represented in F as absorbing states with probability 1. F is thus a proper right stochastic matrix, with each row summing to 1. The aggregate proportion of food f_{uw}^{agg} sent from producing node u in stage $V(1)$ along all possible network paths to w in stage $V(n)$ can be found as:

$$f_{uw}^{agg} = \left[(F_o)(F)^{n-1}\right]_{u,w}. \qquad (2.1)$$

Though we assume that flows are strictly bipartite, both the network model and traceback framework presented here can accommodate flows across multiple stages. Extensions of the method to networks with inter-stage links can also be made.

In sum, the food distribution network $G\{V,E,N\}$ is fully characterized by the normalized initial volume matrix F_0, a square transition probability matrix F composed of the normalized flow proportions f_{uv} between any adjacent nodes u and v, and a matrix of the same dimension D representing the distances d_{uv} between adjacent nodes.

2.1.3. Contamination diffusion model

Key Assumptions

We assume that at the source, contaminated product is subdivided into many individual batches. At time t_s, multiple truckloads conveying batches containing contaminated product will depart from contaminated source node s^*, heading along out-going edges to nodes $u \in V(2)$. Each truck departure entails a direction and transport time chosen independently, such that multiple trucks may travel separately but in the same direction. At infected downstream nodes, batches of contaminated product will be separated and then re-aggregated with other batches, contaminated or not contaminated, into new transportation units. The process will then continue, with the contaminated units being distributed stochastically to downstream nodes. In practice, we can reasonably expect this condition to be validated, since due to the small fraction yet widespread distribution of case reporting, many more items leading to contamination will travel separately than will share batches.

We also assume that the total initial volume of contaminated product is conserved, meaning that once the contamination enters the supply chain it will not spread or grow, a conservative assumption in practice as the concentration of contamination will likely decay during its journey through the supply chain (LeBlanc et al. 2015; McKellar et. al 2014). We emphasize two characteristics that distinguish this model from a contagion model of contamination spreading as a result of this assumption. First, the contamination is not necessarily dispersed to all downstream nodes; the span of downstream nodes receiving contaminated product will depend on the size of the initial contamination and the particular day's logistics. Second, while in the following we refer to any node *v* receiving contaminated product as "contaminated" or "infected," this implies only that at least one batch of contaminated product took a tour through *v* and not that all of the product at node *v* has become contaminated.

Diffusion Model

We can now take the perspective of an individual contaminated item *i* traveling within a single batch as it makes its journey through the supply chain to an eventual illness observation at contamination time t_i.

The diffusion process is initiated by the contaminated source node $s^* \in V(1)$ at an unknown time $t = t_s$. $V(1)$ can represent any (upstream) class of supply chain actors, though is most often a Farm or Processor. We model s^* as a random variable (RV) with a predefined prior probability distribution,

$$P(s^* = s) = \left\{ \begin{array}{ll} p_s & s \in V(1) \end{array} \right\}. \qquad (2.2)$$

In the absence of any prior information external to the distribution network, the prior probabilities are determined from the proportion of food originating at each node $s \in V(1)$, such that $p_s = f_s$.

From s^*, *i* departs for a downstream node $u \in V(2)$, chosen randomly according to transmission probabilities $p_{sv} = f_{sv}$. This process continues across the stages of the supply chain until *i* reaches a retailer node o_i in stage *N*, generating a unique path γ_{si} defined by a set of nodes in each stage, $\gamma_{si} \triangleq \{s, u, \ldots, o_i\}$. Transportation times between supply chain nodes *u* and *v* are drawn from a distribution $D_{uv} = \mu_{uv} + \theta_{uv}$, where $\mu_{u,v}$ is a deterministic component relating to the distance $d_{u,v}$, and $\theta_{u,v}$ is a random delay variable centered on the origin. A second type of random delay is associated with storage at each supply chain node *u* is chosen from a non-zero random variable θ_u. From here, *i* is stored at retail until it is purchased, consumed, and finally, after an incubation period, results in an infection reported at time t_i, according to a patient's recalled time of illness onset. Due to the inaccuracy inherent in this patient-estimated onset time, we model t_i as being distributed uniformly over an uncertainty window equal to $[t_i \pm \tau]$. Finally, a third type of random delay associated with storage at retail, storage at point of consumption, and incubation time, is chosen from a non-zero random variable $\theta_{\alpha\beta}$, parameterized by the type of food α, determining the storage time, and pathogen β, determining the incubation time. Each infection time t_i is associated with the retail node $o_i \in V(N)$ from which the offending product was purchased. The total delay density T_{si} from t_s to illness at t_i is thus distributed as $T_{si} = \sum_{(u,v)\in\gamma_{si}} D_{uv} + \theta_{\alpha\beta} + \theta_u$.

An important feature of our model is that due to the independence of departures from s^*, the direction and time of transmission between nodes are independent, as are the path probabilities and total delay densities. By the central limit theorem, the transmission density T_{si} is well approximated by a Gaussian RV, even for differently distributed delays θ_j, providing these random variables have finite variances. This condition is always met in practice, as all food items have a finite lifetime.

By some time t_W, an observation window cutoff, a set O of K observations of illness will have been recorded and linked to the set of associated retailer nodes $H \subset V(N)$. Each observation $i \in O$ is composed of the time t_i and location o_i of contamination, $o_i \in H$, such that $O \triangleq \{(t_i, o_i)\}_{i=1}^{K}$. Note that the o_i are not necessarily unique and $|H| \leq K$, since a node o_i may be linked to multiple cases of illness reached by independent paths and reported at different times.

2.1.4. Source detection

Our goal is to find the source s^* from the K contamination times $\{t_i\}_{i\in O}$ linked to infected retailer nodes H. We introduce a maximum a posteriori probability criterion $\hat{S}$ that selects the source node $\hat{s} = s$ that maximizes the likelihood of the observations, and the prior distribution for s^*:

$$\hat{s} = \arg\max_{s\in\Omega} P(s^* = s) P(\{t_i\}_{i\in O} | s^* = s), \qquad (2.3)$$

where $s \in \Omega$ is the set of feasible source nodes; that is, the set of nodes in $V(1)$ that have at least one path to all contaminated nodes H. Since the probability of the observation times depends on the unknown start time t_s, we rewrite the likelihood as:

$$P(\{t_i\}_{i\in O} | s) = \max_{t_s} P(\{t_i\}_{i\in O} | s, t_s). \qquad (2.4)$$

The source detection approach is to first find, for each possible source s, the value for t_s that maximizes the probability of the observations $P(\{t_i\}_{i\in O} | s, t_s)$ by varying t_s over feasible times, $t_s \in (-\infty,\ \min\{t_i\}_{i\in O})$. After finding a value for t_s and its associated likelihood $P(\{t_i\}_{i\in O} | s, t_s)$, we choose s with the maximum a posteriori probability. The initiation time is left over, for free! A strategy for optimizing the objective in (4) is developed below.

Approximate objective function

Maximizing t_s corresponds to solving:

$$\hat{t}_s = \max_{t_s} P(\{t_i\}_{i\in O} | s, t_s) = \max_{\pi_s \in \Pi_s} \max_{t_s} P(\{t_i\}_{i\in O} | s, t_s, \pi_s) P(\pi_s | s) \qquad (2.5)$$
,

where $\pi_s \triangleq \{\gamma_{si}\}_{i \in O}$ denotes the collection of paths from s to all observations $i \in O$, which we call a *cascade*, and $\prod_s$ is the set of all possible cascades. By factoring the probability of the cascade π_s out of the start time likelihood, we isolate the problem into a the temporal component and a volume component: The first term to the right of the maximization represents the likelihood that the observation times $\{t_i\}_{i \in O}$ are observed given the cascade π_s, a probability density over time. The second term represents the probability that the particular set of paths in the cascade π_s is taken by contaminants $i \in O$, determined by the transmission probabilities $p_{sv} = f_{sv}$, representing the relative volume of commodity traded.

Equation (2.5) decomposes the start time maximization problem into the sub-problem of computing and comparing the maximum probability initiation time of each possible cascade $\pi_s \in \prod_s$. Due to the combinatorial nature of $\prod_s$, however, the complexity of (2.5) grows exponentially and is therefore intractable. To solve (2.5), we introduce an approximation. First, we assume that the actual diffusion cascade is the maximum probability cascade, π_s^m. Due to the independent path assumption, π_s^m is the collection of maximum probability paths γ_{si}^m from source s to all observation nodes:

$$\gamma_{si}^m = \max_{\gamma_{si} \in \Gamma_{si}} P(\gamma_{si}|s) = \max_{\gamma_{si} \in \Gamma_{si}} \prod_{(u,v) \in \gamma_{s,i}} p_{uv}, \qquad (2.6)$$

where Γ_{si} is the set of all possible paths from s to i. The resulting objective can be written as

$$\hat{t}_s = \max_{t_s} P\left(\{t_i\}_{i \in O} | s, t_s, \pi_s^m\right) P\left(\pi_s^m | s\right). \qquad (2.7)$$

Improved approximate objective function

We now consider the implications of the maximum probability cascade assumption. For *heterogeneous* network structures in which there is considerable variance in both path lengths and the distribution of links and flows, the paths traveled by contaminated product being distributed stochastically through the distribution network will be dominated by the largest flow probabilities (Grady et al. 2011; Brockmann and Helbing 2013). The highest probability paths will accumulate a greater fraction of the overall transmission of contamination, and Equation (2.7) will model the actual paths traveled by a larger fraction of observed contaminants. However many food distribution networks exhibit more *homogeneous* structure, with equal or close to equal path lengths and degree distributions with lower variance. The less differentiated the path probabilities, the likelier it is that contaminated product will take multiple paths from s to i, and the less Equation (2.7) will be capturing the actual paths travelled by the observed contaminants. In the extreme, when multiple paths of equivalent probability exist, Equation (2.7) will still consider only one, and will considerably undercount the total transmission probability of s to i along all possible paths. To avoid this error and design our method for the network structure characteristic to the problem at hand, we improve upon our first approximation by replacing the maximum cascade probability in Equation (2.7) with the aggregate probability $P\left(\pi_s^{agg}|s\right)$, the cumulative probability of flows from s to $i \in O$ along all possible network paths. $P\left(\pi_s^{agg}|s\right)$ is found as

$$P\left(\pi_s^{agg}|s\right) = \prod_{i \in O} f_{si}^{agg},$$

where f_{si}^{agg} is defined in (1). The resulting objective can be written as

$$\hat{t}_s = \max_{t_s} P\left(\{t_i\}_{i \in O} | s, t_s\right) = \max_{t_s} P\left(\{t_i\}_{i \in O} | s, t_s, \pi_s^m\right) P\left(\pi_s^{agg} | s\right), \qquad (2.8)$$

where the first term to the right of the maximization represents the likelihood that the observation times $\{t_i\}_{i \in O}$ are observed given the set of maximum probability paths are traveled to each i, and the second term represents the total transmission probability from s to i along all possible paths.

Initiation time maximization problem

With the approximate objective fully specified, we now develop a solution to the optimization problem,

$$\max_{t_s} P\left(\{t_i\}_{i\in O}\middle|s,t_s,\pi_s^m\right), \qquad (2.9)$$

where the transmission probability component in Equation (2.8) has been left out because it does not depend on time.

The goal of Equation (2.9) is to find the start time that maximizes the observed contamination times, assuming the contaminated items leading to those observations traveled along the highest probability paths, and according to the diffusion model of Section 2.1.2. An important implication of the independence assumption of the underlying diffusion model is that all products departing at the same time from s are modeled as traveling independently from s to i, even if links along that path are shared. In the context of the start time maximization problem, this entails that the time delay density $P\left(t_i\middle|s,t_s,\pi_s^m\right)$ for each observation $i\in O$ are independent and can be considered individually. Thus, we can factorize the likelihood in (2.9) as

$$P\left(\{t_i\}_{i\in O}\middle|s,t_s,\pi_s^m\right)=\prod_{i\in O}P\left(t_i\middle|s,t_s,\gamma_{si}^m\right)$$

where for each infected node, the above term can be further written as

$$P\left(t_i\middle|s,t_s,\gamma_{si}^m\right)=P\left(t_s+T_{si}^m\middle|s,t_s,\gamma_{si}^m\ \in\left[t_i\pm\tau\right]\right)=\int_{t_i-t_s\pm\tau}f\left(T_{si}^m\middle|s,t_s,\gamma_{si}^m\right)dt, \qquad (2.10)$$

given the inherent uncertainty τ around the contamination time. As defined in Section 2.1.2, the propagation delay T_{si}^m follows a Gaussian distribution $N(\mu_{si},\sigma_{si}^2)$ with parameter values

$$\left\{\ \mu_{si}=\sum_{(u,v)\in\gamma_{si}^m}\mu_{uv}+\mu_{\alpha\beta} \qquad \sigma_{si}^2=\sum_{(u,v)\in\gamma_{si}^m}\sigma_{uv}^2+\sigma_{\alpha\beta}^2\right. \ .$$

The resulting maximization can be computed efficiently using line search methods. The optimization will perform best when the variance of these distributions is low, meaning the travel and storage delays are well understood, and when there is significant displacement between their means, meaning each distribution is clearly associated with a particular path traveled.

A graphic interpretation of the start time maximization problem involving three observations is provided in Figure 2.2. For a given candidate source node $s\in\Omega$, we determine the shape of the delay density distribution $P\left(t_i\middle|s,t_s,\pi_s^m\right)$ for the highest probability path to each observation $i\in O=\left\{(t_i,o_i)\right\}_{i=1}^3$. Each delay density distribution is a function of the contamination initiation time t_s. For an assumed value of t_s, we calculate the probability that each observation time t_i, padded by its uncertainty window τ, falls within the delay density distribution. These probabilities are equal to area within $\left[t_i\pm\tau\right]$ falling under each the density curve, as depicted by the shaded region under each curve. One can imagine moving t_s, and with it the delay densities T_{si}^m, forward or backward in time so that the areas in the shaded regions grow or shrink. To compute the joint likelihood across all observations i that node s is the true source, given the assumed value t_s, by independence these probabilities are multiplied, forming $\prod_{i\in O}P\left(t_i\middle|s,t_s,\pi_s^m\right)$. After varying t_s extensively through line search optimization methods, the start time estimate $\hat{t}_s$ is ultimately chosen as the value of t_s that maximizes the joint likelihood. The more peaked the delay densities and the greater the displacement between them; that is, the lower the variance and the greater the difference in their means, the more distinguished the optimal solution will be and the better the resulting estimate.

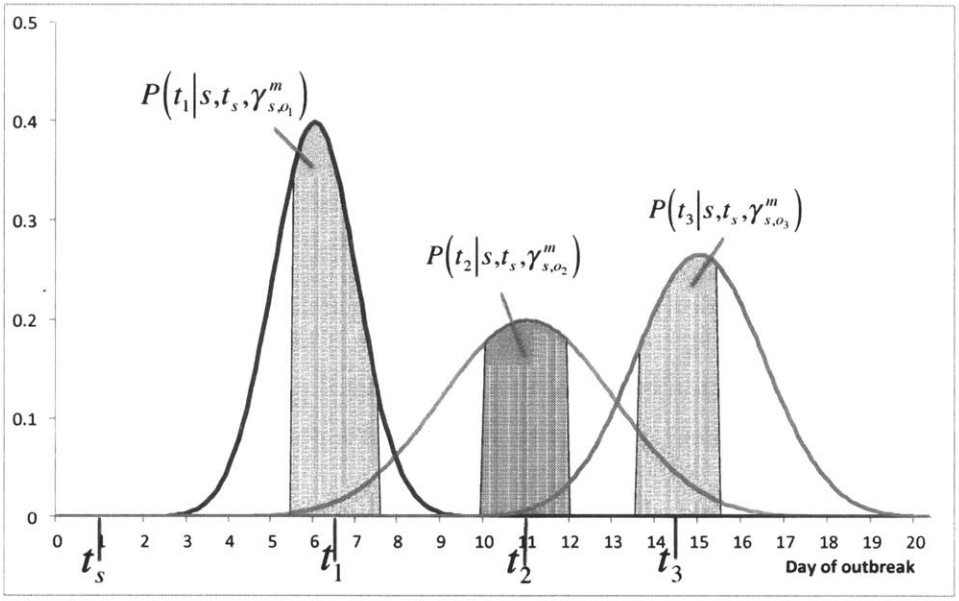


Figure 2.2. Graphic representation of the initiation time maximization problem for 3 observations.

Identifying the most likely source

We now have all the pieces necessary to solve the approximate objective function, which can be written:

$$\hat{s} = \arg\max_{s\in\Omega} P(s) \max_{t_s} P\left(\{t_i\}_{i\in O} \middle| s, t_s, \pi_s{}^{m}\right) P\left(\pi_s^{agg} \middle| s\right), \qquad (2.11)$$

where $P(s)$ is the prior distribution for s^*, the first term to the right of the start time maximization represents the likelihood that the observation times $\{t_i\}_{i\in O}$ are observed given the set of maximum probability paths are traveled to each *i*, and the second term represents the total transmission probability from *s* to *i* along all possible paths. Finally, the posterior probability is determined for source *s* through Bayesian updating the likelihood of the observations given *s* with the prior probability for *s*.

Equation (2.11) chooses the maximum probability source according to the approximate objective function. A posterior probability can be constructed for each feasible source $s\in\Omega$. By normalizing the a posteriori probabilities, we can form a probability mass function (PMF),

$$P\left(s^* = s \middle| \{t_i\}_{i\in O}\right) \qquad (2.12)$$

over the set $s\in\Omega$. The resulting PMF can be used to identify a set of the most probable sources.

We summarize the algorithm in Table 2.1.

2.1.5. Discussion

Key features and contributions

In this section, we presented an original formulation of and solution to the problem of locating the source of an outbreak of foodborne disease. The model and algorithm presented leverages the structural, volume, and temporal dimensions of the problem while accounting for inherent uncertainties. In particular, there are multiple sources of uncertainty in the temporal dimension: the inherent inaccuracy in the times themselves, recorded according to a patient's recalled day of illness onset, and the uncertainties accumulating across each distribution modeling delays in travel or storage through the supply chain. The temporal component will perform best when the variance of these distributions is low, meaning the travel and storage delays are well understood, and when there is significant displacement between their means, meaning each distribution is clearly associated with a particular path traveled. To capture the predictive information contributed by the temporal dimension when these conditions are met while accounting for those cases when they are not, our solution balances the temporal contribution with the spatial contribution, weighting the likelihood of the maximum probability time with the transmission probability to form the total likelihood for *s*.

```
Inputs:
    · $G\{V,E,N\}$, food distribution network with initial volume, flow, and distance matrices $F_0$, $F$, and $D$
    · $\mu_{si}$ and $\sigma_{si}^2$, parameters for the Gaussian propagation delay density $T_{si}^m$, specific to $G$, commodity $\alpha$ and pathogen $\beta$
    · $P(s^* = s)$, prior distribution
    · $O$, illness observation set at time $t_W$
    · $\tau$, contamination time uncertainty
For $u \in V(1)$:
    If $u$ reaches all observed contamination nodes $o_i \in H$, add to feasible source set $s \in \Omega$
For $s \in \Omega$:
  For $o_i \in H$:
    Determine $\gamma_{si}^m$, the maximum probability path from $s$, using Equation (6)
  Find $\hat{t}_s = \max_{t_s} P(\{t_i\}_{i\in O} | s, t_s)$ using Equations (8-9), and line search optimization
Return
    $\hat{s} = \arg\max_{s\in\Omega} P(s) \max_{t_s} P(\{t_i\}_{i\in O} | s, t_s, \pi_s{}^m) P(\pi_s^{agg} | s)$.
```

Table 2.1. Contamination Time Source Detection Algorithm

A further contribution of this solution is its specificity to the problem context, tailored to the variability often characteristic of distribution network structures, without loss of generalizability to cases when differentiation is less pronounced. This is achieved through a second system of balances: the combination of an aggregate probability term in the transmission probability with a maximum probability term in the delay likelihood. The maximum probability term will perform best when there is greatest heterogeneity in the structure and flows. The aggregate probability term will provide a greater advantage in homogeneous networks while not penalizing the heterogeneous case.

The toy example in Figure 2.3 illustrates the method in application and demonstrates its advantages. Figure 2.3a is a completely *homogeneous* network, where all nodes are connected to the same number of nodes, and all flows are equal. Figure 2.3b exhibits the same structure but non-identical flow probabilities and therefore represents a slightly more *heterogeneous* case. We apply the traceback algorithm to predict the source of contamination for the two network structures, given a scenario in which two contaminations have been observed and two feasible sources exist. The example demonstrates that the algorithm performs best when there is variability in the distribution of links and flows, but that even when links and flows are identically distributed as in the *homogeneous* case, our method is able to distinguish between the two sources.

We apply the traceback algorithm to predict the source of contamination for the two network structures. First, we determine the set of feasible sources, the first stage nodes sharing at least one path to both *K* and *O*, to be $\Omega=\{D,E\}$. Now we find the probability that *D* and *E* are the true source, starting with *D*. We see that *D* reaches *K* along one path, $\gamma_{D,K}^{1}=\{D,J,K\}$, and *O* through two paths, $\gamma_{D,O}^{1}=\{D,I,O\}$ and $\gamma_{D,O}^{2}=\{D,J,O\}$. To determine the transmission probability term, we calculate the aggregate probability from *D* to both observations, summing over the probabilities of the three paths $P(\gamma_{D,K}|D), P(\gamma_{D,O}^{1}|D)$, and $P(\gamma_{D,O}^{2}|D)$. In the case of the *homogeneous* network in Figure 3a, the three paths are of equal probability 0.25, for a total aggregated probability $P(\pi_s^{agg}|D)=0.75$. In the case of the more *heterogeneous* network in Figure 3b, the paths are of differing probabilities $P(\gamma_{D,K}|D)=0.45$, $P(\gamma_{D,O}^{1}|D)=0.05$, and $P(\gamma_{D,O}^{2}|D)=0.45$, for a total aggregate of $P(\pi_s^{agg}|D)=0.95$.

To perform the start time maximization, we choose the highest probability path from *D* to each observation *K* and *O*, determine the shape of the delay distribution for each of the two paths, and perform the maximization according to the process in Figure 2.2 to determine $\hat{t}_D=\max_{t_D} P(\{t_K,t_O\}|D,t_D,\pi_s{}^m)$. For the *heterogeneous* network in 2.3b, the path $\gamma_{D,O}^{2}$ is the clear higher probability choice for the algorithm. Since there is a 9 times greater probability that the contamination actually traveled to *O* via this path, there is a correspondingly greater chance that the resulting time maximization will perform well. In contrast, both paths to *O* in the *homogeneous* network are of equal probability and so the algorithm chooses one at random; if the actual path traveled is not the one chosen by the algorithm, the start time estimate will likely result in greater error, though this will ultimately depend on how uncertain and similarly distributed the delay densities are.

To determine the posterior probability $P(s^*=D|\{t_K,t_O\})$ that that *D* is the source, we multiply the probability component from the time maximization, $P(\{t_K,t_O\}|D,\hat{t}_D,\pi_s{}^m)$, with the aggregate path probability $P(\pi_s^{agg}|D)=0.75$, and combine with the prior probability $P(s^*=D)$. We now repeat the process for the other feasible source *E*. *E* connects to both *K* and *O* along only one path each, $\gamma_{E,K}=\{E,J,K\}$ and $\gamma_{E,O}=\{E,J,O\}$, so no differentiation is necessary to perform the start time maximization. The aggregate probability term is the same for both networks, equal to $P(\pi_s^{agg}|E)=0.5$.

Finally, we compare the resulting posterior probabilities to choose between *D* and *E* as our prediction for the true source. Focusing on the contribution of the aggregate probability term to the prediction, *D* is clearly established as the likelier source for both cases, though the result is more pronounced for the *heterogeneous* network, where $P(\pi_s^{agg}|D)=0.95$ > $P(\pi_s^{agg}|E)=0.5$, compared with the *homogeneous* case, where $P(\pi_s^{agg}|D)=0.75$ > $P(\pi_s^{agg}|E)=0.5$. The prediction may be further improved with the probability contribution from the start time maximization, especially in the case of the *heterogeneous* network, though again this will ultimately depend on the uncertainty in the delay densities. Thus, our solution is able to benefit from the signal sent by the variability in the *heterogeneous* case, while still being able to differentiate between the two sources when links and flows are identically distributed.

In the initial evaluation presented in Section 2.3 and the more extensive results in Chapter 3 we will demonstrate this behavior on a larger scale, demonstrating the high performance of our methodology across a variety of distribution network structures.

Critical assumptions

The major assumption of this derivation is that contaminated items leading to observations of illness travel independently through the supply chain. As asserted above, this assumption is reasonably expected to be validated in practice, since the small fraction yet widespread distribution of case reporting observed for large-scale, multi-state foodborne disease outbreaks means that many more items leading to contamination will travel separately than will share batches. Of course, it is possible that food items resulting in contamination might travel in a batch, sharing multiple steps of their tour through the supply chain. In these cases, the optimal solution to the time maximization problem would be found in a two-stage approach, first determining the most likely start time at the most recent shared node, and second determining the start time by optimizing over those interior solutions. Since interior nodes are fewer steps away from the source node, there will be a tighter bound on the uncertainty in the estimate. In this particular problem context, however, optimizing the arrival time at intermediate nodes would contribute only marginal improvement, if any, for three reasons: (i) there is already great uncertainty in the time estimates, both from the inherent inaccuracy in the date recorded, and in the uncertainties accumulating across each network travel and storage delay, (ii) it is not possible to determine, based on the data available, when items leading to contamination have traveled in the same batch, and (iii) it is more likely that items have traveled separately than have shared batches. As a result, we have adopted a solution that better reflects the situation in practice.

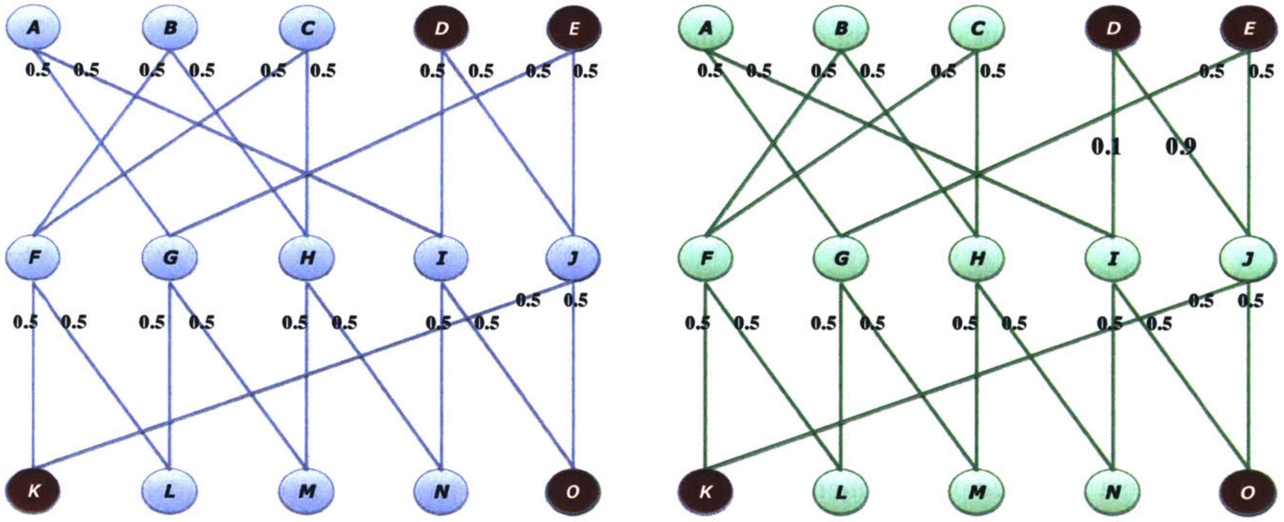


Figure 2.3: Illustration of the traceback method in application and demonstration of its ability to perform well for various network structures. **Figure 2.3a** (left) pictures a completely *homogeneous* network, where all nodes are connected to the same number of nodes, and all flows are equal. **Figure 2.3b** (right) exhibits the same structure but non-identical flow probabilities and therefore represents a slightly more *heterogeneous* case. Both figures depict the same scenario in which node K and O have been contaminated, with one observation each.

2.2. Performance Evaluation Framework

2.2.1. Overview

In the remainder of this chapter, we subject the Spatio-temporal Traceback technique to initial studies to evaluate its performance. First, we develop a simulation-based evaluation framework that allows us to measure the success of the algorithm across a wide range of outbreak scenarios and network structures. Underlying this framework are a network structural model and contamination simulation model. The contamination model is used to generate contamination events in the food distribution network, creating cascades of contamination through a network that eventually lead to reports of illness at specific times and node locations. At a slice in time in the outbreak's progression, the traceback algorithm is applied and a PMF over the feasible sources is constructed using Equation (2.12). The feasible sources are then rank-ordered according to their probability values. To assess the traceback accuracy of this particular network structure, multiple contamination events are generated and the cumulative results assessed using the accuracy metrics described below. This process can be repeated at various intervals as the contamination event progresses and illnesses continue to present, generating a series of rankings as a function of time or case development.

In the following, we summarize the key properties, features, and parameters of the outbreak contamination model (2.2.2) and the network generating model (2.2.3). We then describe the experimental setting and accuracy metrics we use for determining success (2.2.4).

2.2.2. Outbreak contamination simulation model

Outbreak contamination simulation model overview

A Monte Carlo discrete event simulation model was built to generate outbreaks, trajectories of contamination through the supply chain, and reports of illness. The model systematizes and parameterizes each component of the contamination event, spreading, and reporting system. A contamination initiation event is generated according to a set of parameters defining the initial conditions defining an outbreak and its mode of dispersion through the network. Once the contamination event has been generated, a set of probability distributions defines the dynamics of the contamination spreading process, from distribution between and storage at supply chain nodes, to purchase and storage at destination (home or restaurant) before consumption, to disease incubation period after the contaminated product has been consumed, until symptoms present and medical attention is sought. A graphic overview of the contamination spreading process is provided in Figure 2.4. Multiple reports of illness are generated across time, which are assembled into a simulated epidemic curve according to the date of illness onset. The parameterization of the initial conditions allows for a combinatorially large set of possible outbreak scenarios to be generated, while the stochasticity of the dynamical contagion process allows that each epidemic resulting from the same initial conditions to take a different form.

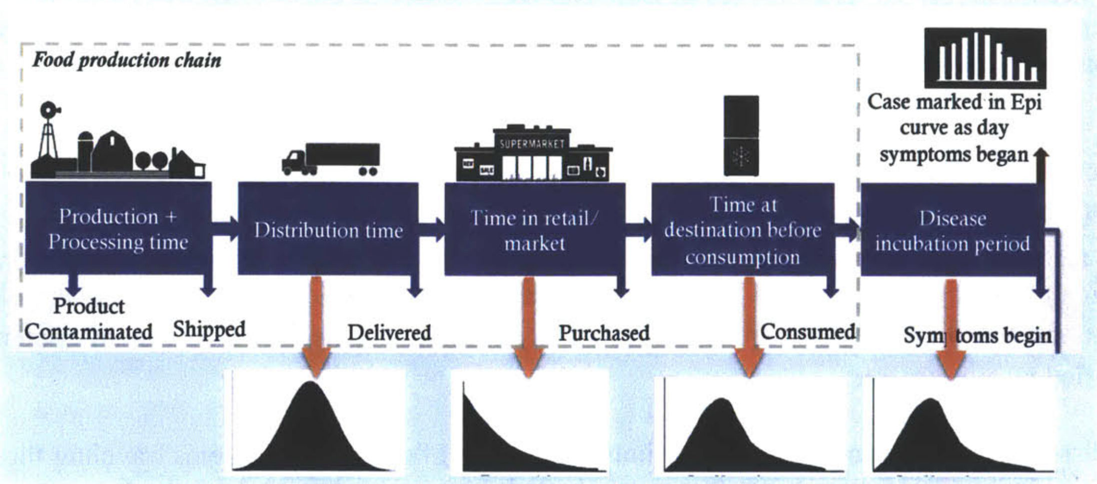


Figure 2.4. Graphic overview of the contamination spreading process. Once a product has been contaminated, a set of probability distributions defines the dynamics of the contamination spreading process, from distribution between and storage at supply chain nodes, to purchase and storage at destination before consumption, to disease incubation period after the contaminated product has been consumed, until symptoms present and medical attention is sought. Multiple reports of illness are generated across time, which are assembled into a simulated epidemic curve according to the date of illness onset.

Data Sources

The distributions and parameters representing supply chain dynamics are informed by models and data in the published literature, overviewed below. Real data and expert elicitation were used to "reality check" and tweak resulting modeling choices. Further details on the specific distributions and parameters provided by these sources are found in Appendix 2.2. The incubation period for foodborne diseases are estimated from the range of values documented by the FDA (FDA 2016).

Existing work presenting stochastic models of supply chain transport and storage dynamics is limited, and the distributions chosen here were derived primarily from two references, one of which is a review presenting models used in four different studies. Because the distributional parameters will vary for different foods, we also reviewed field studies presenting raw data not fit to distributions in order to compare results and estimate ranges for parameter values. The general shape of probability distributions were informed by stochastic models presented in Laguerre et al. (2013) and Poulliot et al. (2010). Laguerre et al. (2013) present a review of stochastic models derived from field studies measuring the time a commodity spends at various stages in the supply chain. The probability distributions reported there are used to inform the distribution for time spent in storage at processor, warehouse, and retail, as well as storage at destination before consumption. Poulliot et al. (2010) derive survival distributions for the time various types of food spend in home storage using consumer survey data. Their distributions were similar to those presented in Laguerre et al. (2013). A review of the existing literature reporting supply chain transport and storage time data recorded in field studies or used in non-parametric simulation models helped to determine ranges of parameters. McKellar et al. (2014) measured residence times in a retail supply chain from a processing facility to retail storage for lettuce in both winter and summer months. Using the storage and delay times reported by McKellar et al. (2014), Hashemi Beni et al. (2011, 2012) and LeBlanc et al. (2015) created a simulation model of the contamination spreading process in the bagged lettuce retail supply chain in Canada. Dallaire et al. present a small case study tracking broccoli through a supply chain involving 2 growers, 1 wholesaler, 4 retailers in same geographical area, measuring time spent in each transit and each storage point. Finally, data previously collected and used in an FDA food safety consequence management system (BT Safety, LLC) and anonymized shipping data shared by a large foodservice distributor and two national lettuce producers were used to "reality-check" distribution and parameter choices (See Appendix 2.2).

Outbreak Contamination Model

First, a contamination event is generated at $t_s^* = 0$. We assume the contamination originates at a single node, and that this node is in stage $n = 1$. The source node s^* is chosen probabilistically, according to the proportion of total initial volume f_s at each node $s \in V(1)$, thus enforcing an assumption that contaminations will originate more frequently for larger volumes handled. The volume of contamination c_s^* is specified by a parameter.

Multiple dispersion modes can be chosen to simulate the path of contaminated items traveling through the supply chain. In *standard* dispersion, the contaminated product is separated into pallets, standardized crates of 40 boxes of produce, which move as independent, identically distributed (IID) units through the supply chain. This mode models even mixing at supply chain nodes, which can reasonably be assumed to be validated in practice, and is thus implemented as the default mode. In minimum dispersion, bulk truck loads carrying 24 pallets are the (also IID) unit of analysis. This mode assumes that pallets are sorted into truck loads at the origin node, and that these loads remain unchanged as they travel through each stage of the network. The maximum number of possible paths forward will be limited to the number of truck loads of contaminated product, which for a day's worth of production from an average size farm, consists of 10 trucks. This mode is therefore meant to model a lower bound on the minimum dispersion of contaminated product. Individual boxes considered as the unit of analysis will model an upper bound on the maximum dispersion of contaminated product.

In any dispersion mode, a unique path through the network and associated time will be sampled for each unit of analysis. Each path will be randomly sampled according to the flow probabilities f_{uv} along the links, until all the capacity along a link has been met. At the final stage of retail nodes, each unit of analysis is disaggregated from bulk load or pallet to individual consumer retail-sized volume.

Following disaggregation and purchase at retail, a path to consumption, contamination, and potential illness is sampled for each individual contaminated produce item:

- Each unit will be purchased at the retailer node.
- Each unit will be consumed by between 1 - 4 people, randomly selected.
- Each person has a probability of developing an illness from the contaminated food according to the virulence of the outbreak strain, called the infectivity, ν .
- Each person developing symptoms has a probability ς of reporting their illness through the medical system.

Once a path through the supply chain has been selected for each unit of analysis, distribution times along links and residence times at nodes are sampled from random variables informed by the literature:

- Transportation times between supply chain nodes u and v are drawn from a Gaussian distribution $D_{uv} \sim N\left(\mu_{u,v}, \sigma_{u,v}^2\right)$, with $\mu_{u,v} = d_{uv}/v_{\text{avg}}$, where v_{avg} is the average velocity of food transport travel, and consider the variance of D_{uv} to be proportional to the mean μ_{uv}, such that $\sigma_{uv}^2 = \left(.5\mu_{uv}\right)^2$. D_{uv} is truncated for some maximum and minimum transport velocity, to ensure that it does not take on infeasible values.

- Storage times at supply chain nodes according to an exponential distribution $S \sim \mathrm{Exp}(\lambda_n)$, where parameter λ_n is specific to the stage n of the node
- Storage time by consumer (e.g. at home) before consumption, also according to an exponential distribution $S \sim \mathrm{Exp}(\lambda_C)$
- Incubation period θ_β for the $\nu \cdot \varsigma$ consumers that develop and report an illness, whose functional form will vary by pathogen type β

Since fresh produce has a limited shelf life, any produce item not consumed within a specified shelf life τ_α for commodity α is discarded. Shelf life begins ticking from the moment of departure from the source node.

Illnesses $i \in O$ are reported at a time t_i and node o_i of contamination, $o_i \in H$, forming a set of ordered pairs $O \triangleq \{(t_i, o_i)\}_{i=1}^{K}$, as defined in section 2.1.2. The outbreak model generates full trajectories of pathways travelled by each contaminated item and can therefore also be used to extract other results such as time of consumption of contaminated food and time of contamination at intermediary nodes along the pathway.

Outbreak scenario baseline parameter specifications

In all analyses presented in this chapter, outbreaks are generated according to the following "baseline" scenario. The parameter (or distributional) specifications summarized in Table 2.2. We note that we choose *E.coli* for the baseline scenario because the incubation period ranges over a longer timeframe than other common foodborne diseases (*Salmonella, Listeria*), thus modeling an upper bound on the amount of uncertainty introduced by this factor.

Baseline outbreak contamination scenario parameter specifications		
c_s^*	Initial contamination volume	1 day of production volume
___	Dispersion mode	"Average"
β	Pathogen	E. coli
ν	Infectivity	1/50
ς	Reporting rate	1/25
v_{avg}	Average transport speed	60 miles per hour
$v_{\max}$	Maximum transport speed	100 miles per hour
$v_{\min}$	Minimum transport speed	10 miles per hour
λ_n	Parameter for storage time at supply chain nodes, for nodes in stage n	$1/\lambda_n = 1$ day for all stages n
λ_C	Parameter for storage time after purchase before consumption	$1/\lambda_C = 3$ days
θ_β	Incubation period distribution, for pathogen type β	$\theta_\beta \sim \mathrm{Weibull}(\lambda, \kappa)$ with scale parameter $\lambda = 4$ and shape parameter $\kappa = 1.5$ (E. coli)
τ_α	Shelf life	30 days

Table 2.2. Baseline outbreak contamination scenario parameter specifications.

2.2.3. Network generating model

As detailed in the introduction, data on structures of aggregated distribution networks is not publicly available. Furthermore, a modeling framework capable of representing these structures does not currently exist. Through a review of the (limited) research efforts documenting specific examples of aggregated distribution structure (Section 2.1.1 and Section 3.1), we are able to learn about the relevant features and parameters of these network structures. We find that these networks have particular structural properties that are not fit by standard network generating models (scale-free, small-world, ERGM, etc.). Therefore we develop a foodborne distribution network modeling framework. Our goal in designing this framework is two-fold: (1) to create a parameterized modeling framework that allows us to perform specific sensitivity and scenario analyses that explore the dependence of traceback accuracy on network structure; and (2) to represent the essential components of this structure, and with enough accuracy to be able to realistically evaluate our traceback approach.

In order to achieve our two goals, we develop two network generating models. Random Layered Graph (RLG) is a completely stylized generating model that represents the specific multi-partite structure found in food distribution networks, parameterizing many of the variables characterizing these particular structures. This model incorporates the structural features of distribution networks but is not otherwise informed by real network data. Regional Network (RN), the second class of model, realistically represents structural, spatial, and temporal properties of foodborne disease networks in the US. This generating model is informed by data on network structure, transport times, and locations and volumes of food production in the US for various food types, and is used to gain insight into the traceback accuracy achievable across these foods.

This chapter focuses solely on results from analysis of stylized network structures generated using RLG, which was designed for the purpose of performing specific traceback accuracy sensitivity analyses.

Summary of network structural model features

RLG exhibits the same essential food distribution network structural features described in the Section 2.1.1, so we adopt the same notational conventions: Food distribution networks are directed, acyclic, *N*-partite graphs $G\{V,E,N\}$ which can be represented by a normalized initial volume matrix F_0, a square transition probability matrix F composed of the normalized flow proportions f_{uv} between any adjacent nodes u and v, and a matrix of the same dimension D representing the distances d_{uv} between adjacent nodes. F_0 can be seen as the initial condition defining the volume of flux along the structural matrix F.

RLG defines and parameterizes multiple variables to specify the structure and distribution of the resulting structural matrix F and initial condition matrix F_0. In the sensitivity analyses presented in this Chapter, we focus on the effect on detection performance of varying the variables with the ranges summarized in Table 2.3.

Network structural variable and distributions varied in Chapter 3 analyses

Name	Variable / Parameter Description	Range of values or distributions varied in Chapter 3
Structural Variables determining F		
n	Number of stages	$n \in [2,6]$
v_n	Number of nodes in stage n	$v_n \in \{10,25,50,75,100,150\}$
X_{out}	Distribution of out-degree links for nodes in stage n	$X_{\text{out}} = \mu$ (*Deterministic*) or $X_{\text{out}} \sim \text{Geom}\left(\frac{1}{\mu}\right)$ (*Geometric)*

μ	Average degree for out-degree distribution X_{out}	$\mu \in [2,10]$
___	Distribution of flow volume across outgoing links from a node	Flow volumes are equal
Initial Conditions determining F_o		
$F_o(s)$	Distribution of initial volume across first stage nodes v_1	$F_o(s) = \frac{1}{v_1}$ *(Deterministic)* or $F_o(s) \sim \text{Geom}(1/\lambda_F v_1)$ *(Geometric)*
λ_F	Parameter governing the spread of $F_o(s)$	$\lambda_F = 5$

Table 2.3. Network structural variable and distributions determining the structural matrix F and the initial conditions F_o . The ranges of values or distributions varied in the robustness analyses of Chapter 3 are presented.

Heterogenity in network structures

Importantly, we are interested in the role of heterogeneity in network structure, which can be modeled by (i) the distribution of the number of links leaving each node across all nodes in a stage, or the out-degree distribution, and (ii) the distribution of flow volumes across the links leaving a single node, or the flow distribution. In this Chapter, we will often make comparisons between networks with and without variability in the out-degree distribution, but with otherwise identical in parameter specifications, referring to these as the *High Variance* and the *Zero Variance* networks. All nodes in the *Zero Variance* network are linked to exactly the same number of other nodes, while in the *High Variance* network the number of out-degrees will vary stochastically. An example of *Zero Variance* and *High Variance* networks that exhibit otherwise identical distributions and parameter values is pictured in Figure 2.5. The distribution of flow volume across outgoing links from a node is not studied in this chapter.

Heterogeneity may also be introduced in the volume distribution $F_o(s)$ across first stage nodes v_1 . We study the effect of heterogeneity introduced in the initial conditions in a specific sensitivity test in Section 2.3.2. In all other studies, we define an initial volume distribution that is equal over the node set $V(1)$ in order to isolate the effect of network structural parameters on detection performance.

Node location assignment

In the analyses presented in this chapter, the matrix D is generated according to the same (random) process for each network. The essential idea behind the distance assignment is to allocate nodes to locations across a rectangular grid according to a set of simple assumptions about the geographical structure of food distribution networks: nodes in the first and last stages are located at random across the grid, and nodes in interior stages are located at the midpoint of incoming or outgoing nodes. The rectangular grid is 2500 x 1500 units, roughly similar in mileage to a rectangle inscribing the United States. Distances are determined by computing the spatial distance between adjacent nodes. This result of this assignment is a distribution pattern in which interior nodes cluster in the center of the grid, and nodes in the first and last stages are distributed equidistantly, on average, from the center. Because all nodes of the same type are distributed similarly within the grid, all network paths from source to contamination node are, on average, of a similar total distance. This in effect means that for networks generated by RLG, the geographic distribution of nodes will have little to no role in detection performance. We study the role of heterogeneity in geographic distribution on detection performance in Chapter 4 with the RN model.

2.2.4. Implementing the Algorithm

With the key properties, features, and parameters of the outbreak contamination simulation model and the network generating model developed, we now describe the experimental setting and accuracy metrics we use for evaluating the performance of the traceback methodology.

Experimental setting overview

Simulation experiments are performed to determine the accuracy of the traceback algorithm applied to a network $G\{V,E,N\}$ with structural, initial condition, and distance matrices F, F_o, and D and a specific outbreak contamination scenario.

In each experiment, we generate 100 contamination events from randomly selected outbreak sources. Each contamination event generates an outbreak source s^* and a set of illness observations O as a function of time and location. The set O can be sorted into intervals of size ω across multiple dimensions: time observation window t_ω, number of illnesses K_ω, or number of contaminated nodes H_ω, forming sets O_ω. An epidemic curve E can be formed by plotting the cases O_ω according to their frequency of occurrence. At a desired interval I_W the traceback algorithm is applied.

We then evaluate the performance of the algorithm applied to these 100 outbreak events, at equivalent increments ω. The source localization performance is quantified according to two metrics: Traceback Accuracy and Rank of True Source. The accuracy of the start time estimation is also assessed.

Accuracy Metrics

Traceback Accuracy (TA)

Traceback Accuracy is defined in this study as the percentage of outbreak events for which the true source is correctly identified; that is, for which $\hat{s} = s^*$. TA equal to 1 indicates perfect performance.

Rank of True Source

The traceback algorithm returns a PMF $P\left(s^* = s \middle| \{t_i\}_{i\in O_\omega}\right)$ over the set of feasible sources $s \in \Omega$. The feasible set is rank-ordered according to their probability values, such that the most likely source is ranked in first position. The Rank of True Source metric records the position of the true source node s^* within the ranking, averaging across all outbreak events.

Start Time Estimation Error

Since each simulated outbreak begins at $t_s^* = 0$, the mean of the absolute value of the start time estimate, $E\left[\hat{t}_s^2\right]$, is used to characterize the error of the start time estimation.

Traceback system parameters

The traceback algorithm system parameters used in all studies presented in this chapter are summarized in Table 2.4.

In addition to the network G and the contamination set O_ω, the algorithm requires as inputs:

- μ_{si} and σ^2_{si}, parameters for the Gaussian propagation delay density T^m_{si}, which are specific to network G
- the uncertainty in the contamination time τ, and
- the predefined prior distribution $P(s^* = s)$.

The parameter μ_{si} is further decomposed as $\mu_{si} = \sum_{(u,v)\in\gamma^m_{si}} \mu_{u,v} + \mu_{\alpha\beta}$, where μ_{uv} represents the mean of the travel time from node u to v. We approximate the mean $\mu_{u,v}$ as d_{uv}/v_{avg} where d_{uv} is the geographical distance between *u* and *v*, and v_{avg} is the average velocity of food transport travel, which we set to 60 miles per hour.

$\mu_{\alpha\beta}$ is the mean of the total delay density associated with storage at each specific supply chain node, storage after purchase and before consumption (e.g. at the home), and incubation period, for commodity α and pathogen β. To reflect the process modeled in baseline outbreak scenario where α is spinach and β is E.coli, the parameter $\mu_{\alpha\beta}$ is modeled as 10 days, the sum of the means of each component as informed by the literature in Section 2.2.2. This choice of parameter assumes that the delay distribution is equivalent across all supply chain nodes within a given stage *n*.

The variance of T_{si}^{m} similarly decomposes as $\sigma_{si}^{2}=\sum_{(u,v)\in\gamma_{si}^{m}}\sigma_{u,v}^{2}+\sigma_{\alpha\beta}^{2}$. We consider both terms to be proportional to the mean by the same ratio of spread, such that $\sigma_{si}^{2}=\sum_{(u,v)\in\gamma_{si}^{m}}\left(\frac{1}{2}\mu_{uv}\right)^{2}+\left(\frac{1}{2}\mu_{\alpha\beta}\right)^{2}$.

We set τ equal to 1 day, but note it could be fixed otherwise; we performed many sensitivity tests to confirm results are robust to this value. We model the predefined prior probability $P\left(s^{*}=s\right)$ as being determined from the proportion of food originating at each node $s\in V(1)$, $F_{o}(s)$, assuming that without further information, each food item is equally likely to get contaminated *a priori*. We note that this assumption mimics the contamination model, which samples sources s^{*} according to the initial volume distribution. Here, it is important to point out that in addition to the a priori equality assumption above, we are choosing the prior distribution in this way to deliberately enforce overfitting of the traceback system to the contamination model; we will address our rational in Section 2.3.3, when we consider the sensitivity of the algorithm to the choice of the prior. Furthermore, whenever we define an initial volume distribution that is equal over the node set $V(1)$, this overfitting will not apply.

Traceback algorithm system parameters for studies in Chapter 2		
v_{avg}	Average transport speed	60 miles per hour
$\mu_{\alpha\beta}$	Mean of storage time delay at supply chain nodes, after purchase before consumption, and incubation period	10 days
τ	Contamination time uncertainty	1 day
$P\left(s^{*}=s\right)$	Prior probability	$P\left(s^{*}=s\right)=F_{o}(s)$

Table 2.4. Traceback algorithm system parameters for studies in Chapter 2.

2.2.5. Summary

In this section, we have developed a simulation-based evaluation framework that will allow us to measure the success of the algorithm across a wide range of outbreak scenarios and network structures. The network generating model Random Layered Graph (RLG) is used to create stylized, multi-partite network structures according to a set of parameter specifications. A probabilistic outbreak simulation model generates outbreak cascades in the food distribution network that eventually lead to reports of illness at specific times and node locations, according to a set of parameters governing the contamination event and spreading process. To assess the detection performance of a particular network structure, multiple outbreak cascades are generated and the cumulative results assessed using the metrics Traceback Accuracy and Rank of True Source. In the following sections, this parameterized framework will be used to perform specific sensitivity and scenario analyses that explore the dependence of traceback accuracy on network structure and outbreak parameters. Of the greatest importance is the ability to generate networks for which the structure, determined by the number of links leaving nodes and the distribution of flows across the links, and the initial condition imposed by the volume distribution across first stage nodes, are either fixed to equality or determined stochastically; this will allow us to identify the role of specific types of heterogeneity on detection performance. We are also interested in generating outbreak cascades that result in epidemic characteristics that vary substantially. In Section 2.3, we present results that focus on the specific role of variability in the relationship between detection performance and network topology, and sensitivity to the prior distribution.

The case of simulation

Here, we comment on the features shared by the traceback methodology and the simulation case in order to underscore important differences. Beyond the prior distribution, discussed above, the information essential to the traceback methodology and assumed to be available to investigators is structural: the topology and geography of the distribution network. There are, however, many features describing the dynamics of the outbreak contamination process that cannot be assumed to be known, let alone observable, to investigators at the time of an outbreak, and are thus not explicitly accounted for in the traceback methodology. This applies to the parameters dictating the initial conditions of the outbreak process: the initial contamination volume, dispersion mode, infectivity, and reporting rate. Still, some values are shared by the traceback methodology and the simulation model (the mean of the delay densities), and a degree of overfitting of the traceback algorithm to the simulation case may be occurring. Thus, while in the following Section we will demonstrate that our methodology performs very well for stylized networks and outbreak simulation cases, it will ultimately be necessary to demonstrate its performance (i) when applied to network structures based on real data, as we shall demonstrate using the Regional Network Model, and (ii) when applied to historical outbreak cases, as shall be discussed in Chapter 5.

2.3. Applicability of Method: Numerical Examples

This section provides a first evaluation of the performance of the traceback methodology. We present illustrative numerical examples that demonstrate the applicability of the method and serve as an initial validation of expected or intuitive properties. We also evaluate the sensitivity of traceback performance to the prior probability distribution, demonstrating an important result: that the traceback algorithm converges to the same answer regardless of the initial conditions. This initial evaluation provides insights and implications for practice. Furthermore, it provides a first step in understanding how the accuracy of detection depends on the structure of a network and on the stochastic evolution of the disease trajectory.

2.3.1. Outbreak scenario and network structures

We consider two small yet dense network structures pictured in Figure 2.5: network constructed with out-degree distribution demonstrating *Zero Variance* and *High Variance*. The networks are otherwise constructed according to the identical variable and distribution specifications summarized in Table 2.5. All nodes in the *Zero Variance* network are connected to exactly 4 other nodes, while the number of out-degrees for each node in the *High Variance* is determined stochastically, according to a geometric PMF with distribution $X_{\text{out}} \sim \text{Geom}\left(\frac{1}{4}\right)$, truncated to reflect the fact that any node can connect to maximally as many nodes are in the subsequent stage. The number of nodes is identical for each stage, equal to $v_n = \mathbf{25}$; that is, these are *square layered* structures. For the studies in Section 2.3.2, the initial volume distribution is equal over the node set $V(1)$. For the analysis presented in Section 2.3.3., the initial volume distribution is determined stochastically, sampling volumes from a geometric PMF with distribution

$F_o(s) \sim \text{Geom}(1/\lambda_F v_1)$ where the spread parameter is equal to $\lambda_F = 5$. Detection performance is based on 100 outbreak simulations generated according to the baseline scenario specifications. For specificity, in the following we refer to the two networks as *Zero Variance-4* and *High Variance-4*.

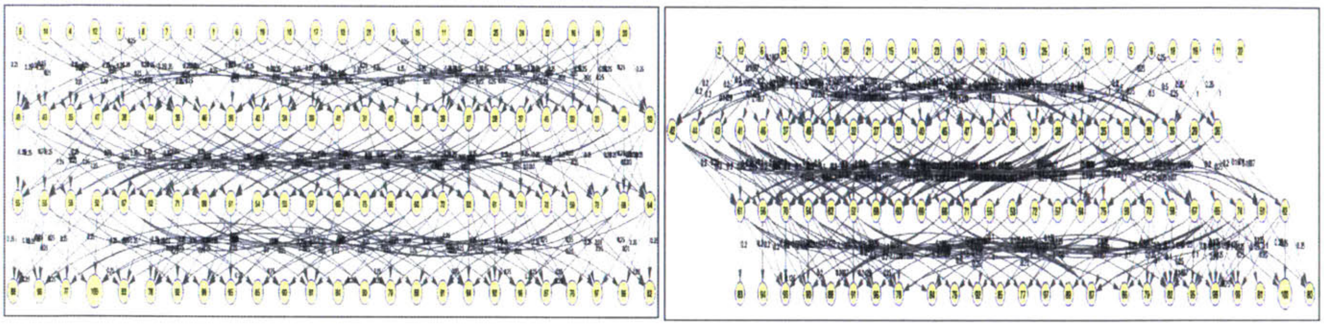


Figure 2.5: Networks analyzed in the numerical example. **(a)** (left) *Zero Variance-4* network, generated with a out-degree for all nodes equal to 4, and **(b)** (right) *High Variance-4* network, generated with the geometric out-degree distribution $X_{\text{out}} \sim \text{Geom}\left(\tfrac{1}{4}\right)$. The networks were generated from the Random Layered Graph (RLG) generating model according to the parameters summarized in Table 2.5.

Structural variable and distributions defining networks studied in Section 2.3.		
Name	Variable / Parameter Description	Range of values or distributions varied in Section 2.3.
Structural Variables determining F		
n	Number of stages	$n = 4$
v_n	Number of nodes in stage n	$v_n = 25$, for all stages
X_{out}	Distribution of out-degree links for nodes in stage n	$X_{\text{out}} = 4$ (*Zero Variance*) and $X_{\text{out}} \sim \text{Geom}\left(\tfrac{1}{4}\right)$ (*High Variance)*
μ	Average degree for out-degree distribution X_{out}	$\mu = 4$
Initial Conditions determining F_o		
$F_o(s)$	Distribution of initial volume across first stage nodes v_1	$F_o(s) = \frac{1}{25}$ ***(Section 2.3.2)*** $F_o(s) \sim \text{Geom}(1/\lambda_F v_1)$ ***(Section 2.3.3)***
λ_F	Parameter governing the spread of $F_o(s)$	$\lambda_F = 5$

Table 2.5. Structural variable and distributions determining the structural matrix F and the initial conditions determining $\boldsymbol{F_o}$, for the *Zero Variance -4* and *High Variance-4* network studied in Section 2.3.2 and 2.3.3.

2.3.2. First Evaluation of Detection Performance

In this section, we demonstrate a first, comparative evaluation of the source detection methodology, investigating the detection performance as a function of case development and the heterogeneity in link connectivity, and the variance in output rankings. We report on (i) the Traceback Accuracy (TA) and the estimation error of the start time $\hat{t}_s$ as a function of $\boldsymbol{K_\omega}$, (ii) TA as a function of the time interval $\boldsymbol{t_\omega}$ and the number of contaminated nodes H_ω, and (iii) the distribution of results for Rank of True source at four contamination intervals $\boldsymbol{K_\omega}$. We compare the performance measures across the *Zero Variance* and the *High Variance* networks described above.

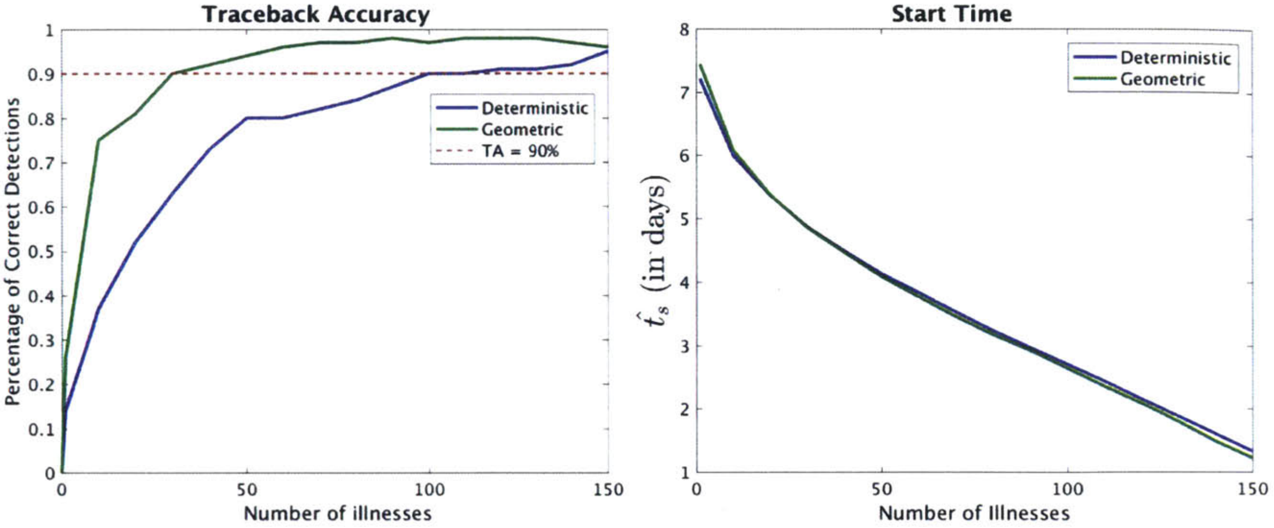


Figure 2.6: (a) Traceback Accuracy (TA) and **(b)** Estimation Error of the Start Time $\hat{t}_s$ as a function of the number of illnesses K_ω, for the *Zero Variance* (blue) and *High Variance* (green) networks. Results are based on 100 simulations in the baseline scenario.

Source detection performs well, especially for networks with greater variance in link distribution
Figure 4 plots Traceback Accuracy (TA) and the estimation error of the start time $\hat{t}_s$ for *Zero Variance-4* and *High Variance-4* as a function of the number of illnesses K_ω. The horizontal dashed line indicates TA equal to 90%. The algorithm clearly performs well; though results vary slightly, accuracy as high as 96% is reached in both cases. The error in the estimation of the start time $\hat{t}_s$ is almost identical for both networks, decreasing almost linearly with the number of illnesses to as low as 1.3 days by 150.

As an initial validation of detection accuracy, these results demonstrate that we can make very good inferences after only a limited number of illnesses have been reported, and almost-perfect inferences if we wait a bit longer. Furthermore, we observe that the detection performance follows expected properties: as the amount of case report data increases, TA increases. Intuitively, something is happening both structurally and statistically to improve accuracy as the reports of illness accrue. Structurally, each node added to the contamination set O_ω acts as an additional topological constraint, shrinking the set of source node suspects Ω that can reach all nodes in O_ω and thus the possibility of error. Statistically, as the number of illnesses K_ω (at not necessarily unique nodes) grows, an increasingly reliable proportion of the contaminated product causing those illnesses will have traveled along the highest probability paths leaving the source node; in other words, the signal sent by the source node will become clearer.

We now comment on differences in the detection performance between the two networks. The accuracy increases very quickly between 0 and 50 illnesses for both networks then begins leveling off at different rates, ultimately converging to peak accuracy at different times. TA improves noticeably faster for the *High Variance-4* network. As demonstrated by the dashed red line in Figure 2.6 a, the algorithm reaches 90% accuracy after 30 illnesses, and converges to peak accuracy of around 96% by 60 illnesses. A more slow and steady increase in accuracy is observed for the *Zero Variance-4* network, which reaches 90% after 100 illnesses and peaks at 96% accuracy at 150 illnesses.

The superior detection performance observed for the *High Variance-4* network follows our expectations. Due to the variance in the distribution of links, the flow probabilities along the links will be more differentiated in this network than in the *Zero Variance* case. Links with larger flow probabilities will dominate the paths traveled by contaminated product through the network, accumulating a greater fraction of the overall transmission of contamination. Recall that the algorithm considers the aggregate probability of travel between each feasible source and contamination node, choosing the source that maximizes the aggregate probability. If these probabilities are more differentiated, the algorithm will be modeling the actual paths traveled by a larger fraction of observed contaminants.

By similar reasoning, we might initially expect the start time estimate to be better for the *High Variance* network. Since the start time likelihood objective is a function of the distance between a feasible source and each node in the contaminated set along the maximum probability path only (rather than the aggregate path probability), it would seem to have even more of an advantage when the maximum probability path is clearly differentiated. These intuitions are sound; rather, the undifferentiated performance that we observe is explained by the distance assignment in RLG. As described in Section 2.2.3, this assignment results in all network paths from source to contamination node being, on average, of a similar total distance. For this reason, the start time likelihood objective contributes little to the algorithm's detection performance in the case of any network generated by the RLG model, though it does result in a high-accuracy estimate of the start time itself.

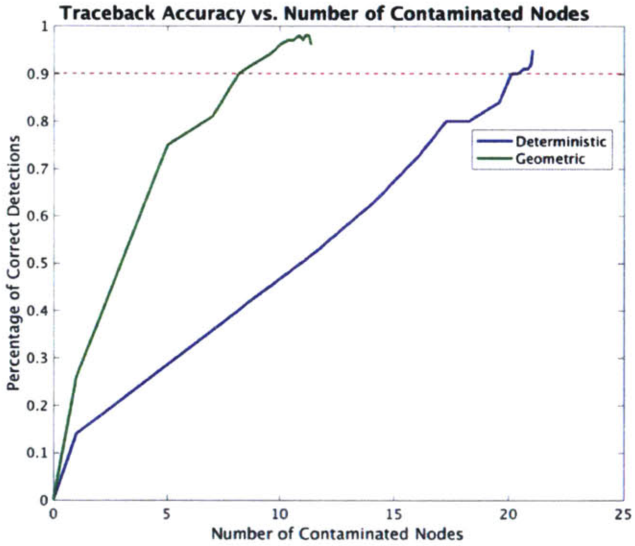


Figure 2.7: Traceback accuracy as a function of the number of contaminated nodes H_{ω}, demonstrating a strong linear relationship within the interval of 0 to around 17 nodes and 5 contaminated nodes, for the *Zero Variance-4* (blue) and *High Variance-4* (green) networks, respectively. The horizontal dashed line indicates TA equal to 90%.

Traceback Accuracy depends strongly on the number of contaminated nodes

Figure 2.7 presents TA as a function of the number of contaminated nodes H_{ω}, for the *Zero Variance-4* and *High Variance-4* networks. As above, the horizontal dashed line indicates TA equal to 90%. In line with the observations above, we note that TA reaches the same peak value for both networks despite many fewer nodes being contaminated, on average, in the case of the *High Variance* network.

We present these results to compare the dependence of detection performance on the number of contaminated nodes with the dependence on the number of illnesses, as in Figure 2.6a. We observe that a clear linear relationship is exhibited for both networks between TA and the number of contaminated nodes within the interval of 0 to around 17 nodes and 5 contaminated nodes, for *Zero Variance-4* and

High Variance-4, respectively. The linear relationship demonstrates that detection performance depends very strongly on the number of contaminated nodes. In comparison to the relationship between detection performance and the number of illnesses, which exhibits a relationship that rises very quickly then levels off, the dependence on the number of contaminated nodes is much more direct, within the respective linear intervals.

It is also important to analyze the relationship between detection performance and number of contaminated nodes beyond the linear interval. Here, the accuracy continues to improve, but at a less regular rate, behavior that is in part attributable to "experimental" causes: there are fewer contamination simulations for which the same number H_ω of nodes were eventually reached. More specifically, we note that in the case of the *Zero Variance-4* network, the TA curve increases directly vertically to reach its end point at 96%. This vertical increase demonstrates that even after the maximum number of nodes has been reached, the accuracy may continue to improve as additional cases accrue at already contaminated nodes. This is also demonstrated by the monotonic increase of the TA curve with the number of illnesses in Figure 2.6a.

Both of these observations: that detection performance depends strongly on the number of contaminated nodes, but that it may still increase after the last contaminated node has been reached, are attributable to the same structural and statistical reasoning provided above. The strong dependence on the number of contaminated nodes further demonstrates the important topological constraint enforced by the contamination set O_ω on the feasible source set Ω, the second shrinking as the first grows. The continued improvement in accuracy after the last node has been contaminated demonstrates that illnesses occurring at already contaminated nodes will work to improve the probabilistic signal sent by the source node.

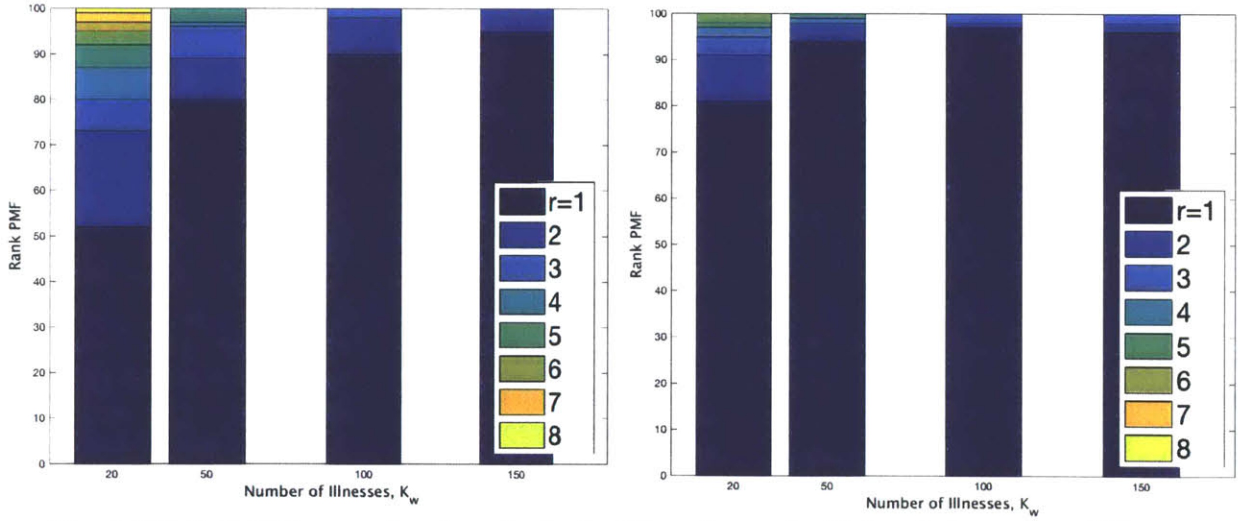


Figure 2.8: Probability mass functions for Rank of True Source *R*, at intervals in number of illnesses $H_\omega = 20,50,100,150$ for **(a)** *Zero Variance-4* network, and **(b)** *High Variance-4* network. Each column demonstrates the Rank PMF at an interval H_ω, where the ascending colors demonstrate the frequency of contamination events leading to traceback performance with rank $r = 1,2,..r_{\max}$.

Variability in detection performance is low when accuracy is high and can be quantified
We now examine the Rank of the True Source metric in order to investigate the variability of the results and quantify their specificity. First, we note that we can interpret the simulation results for Rank as the result of a random variable, since it depends on the random location of the observations. Figure 2.8 plots

the Rank as a random variable *R*, at intervals in number of illnesses $H_\omega = 20,50,100,150$, for (a) *Zero Variance-4* and (b) *High Variance-4*. Each column demonstrates the Rank PMF at an interval H_ω. The ascending colors demonstrate the frequency of contamination events leading to traceback performance with rank $r = 1,2,..r_{\max}$, where $r_{\max}$ is the maximum value taken by *r*.

The frequency plots help us to gain insight into the statistical variability of detection performance in the simulation results. Already at 20 observations of illness, the bulk of the probability density is peaked at $r = 1$ for both networks, though it extends to $r_{\max} = 9$ for *Zero Variance-4* and $r_{\max} = 6$ for *High Variance - 4*. The maximum rank decreases considerably as the intervals increase, and in the case of the *Zero Variance* network, $r_{\max} = 4$ by and $H_\omega = 50$, and $r_{\max} = 2$ by 150 illnesses. We can therefore say with a high degree of certainty that, for this network, the true source will be within the top 5 predictions generated by the algorithm after 50 illnesses have reported, even though the peak percentage of correct detections has yet to be reached. We can also say with a high degree of certainty that if we wait for an additional 100 illnesses, the true source will be within the top 2 predictions. We conclude, therefore, that the *specificity* of the ranking assignment can be quantified to a well-bounded number of possible sources, and that when the Traceback Accuracy is high, this number is low. This is useful information for investigators, as it presents an important tradeoff that we will consider in Chapter 4.

2.3.3. Sensitivity to Prior Probability Distribution

We now evaluate the sensitivity of traceback performance to the prior probability distribution $P(s^* = s)$. Recall that we model the prior probability as being equal to the proportion of food $F_o(s)$ originating at each node $s \in V(1)$, i.e. $P(s^* = s) = F_o(s)$ (see Section 2.2.3). This modeling choice mimics the contamination simulation model, which samples sources s^* according to the initial volume distribution.

In the following, we compare the accuracy of the algorithm with (i) the prior probability implemented as described and (ii) a uniform prior over the set of sources $s \in V(1)$, i.e. any node in stage $n = 1$ is attributed an equal *a priori* likelihood of being the source. To evaluate the impact of the prior distribution on the algorithm's performance, it will therefore be necessary to define an initial volume distribution $F_o(s)$ that is not uniform across the set $V(1)$. For this purpose, we implement the algorithm with (i) and (ii) as described to the *High Variance-4* network structure combined with a geometric initial volume distribution $F_o(s) \sim \mathrm{Geom}(1/\lambda_F v_1)$ as indicated in Table 2.5. We again apply the performance evaluation framework introduced in Section 2.2.

Figure 2.9 plots the Traceback Accuracy achieved with the prior probability set equal to the initial volume distribution ("Prior," indicated by the dotted line), and a uniform prior ("No Prior," indicated by the solid line). Since the prior distribution represents the actual outbreak generating behavior, we would expect its inclusion in the Spatio-Temporal Traceback (STT) algorithm to improve detection performance. The results are surprising. Despite the fact that the prior distribution models the way an outbreak is generated, the algorithm converges to the same answer whether or not this information is incorporated into the model. In other words, the convergence occurs in a way that is agnostic to the choice of prior.

That the algorithm's performance is comparable whether or not proxies to intuitive properties are included demonstrates that the convergence process is dominated by the other dimensions influencing source prediction: network structure, volume, and time. In extreme cases, for example if 90% of volume is produced at one farm node with the remaining 10% distributed across all other farm nodes, we could expect to see a greater influence exerted by the prior, with accuracy much higher. However for situations of the type implemented here, we observe that role of the network structure, volume, and time are more important than are the initial conditions in determining the pathways traveled by contaminated products. This insight has an important practical implication, which is that the initial volume distribution over sources is not a necessary input to the traceback algorithm. In other words, the traceback methodology can be implemented with less information to achieve the same accuracy.

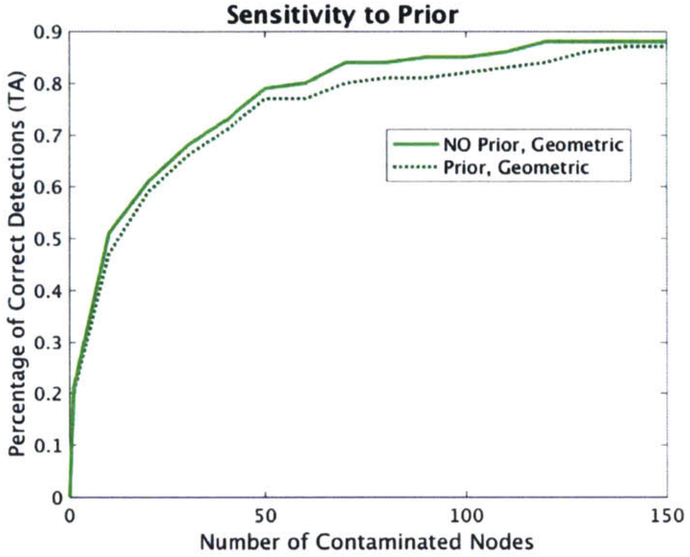


Figure 2.9 Sensitivity of traceback accuracy to prior distribution.

2.3.4. Conclusions

This section presents illustrative performance results for the source detection methodology applied to networks with *Zero Variance* and a *High Variance* in the distribution of link connectivity, but with otherwise identical parameter specifications. Considering these two networks, we have derived four important insights into the detection performance of our algorithm:

1. Traceback Accuracy performs well and follows expected properties, increasing with data both on the number of contaminated nodes and number of illnesses at not necessarily unique nodes;
2. Detection performance is superior for networks with greater variance in link distribution;
3. Variability in detection performance is low when accuracy is high, i.e., the ranking assignment is *specific* to a well-bounded number of possible sources, which can be quantified by the Rank PMF.
4. The algorithm's convergence process depends strongly on the network structure and much less strongly on initial conditions, i.e., on the choice of prior distribution over the sources.is independent of the initial conditions defining the outbreak contamination event.

These insights serve as an initial validation of the behavior and performance of the source detection methodology, and demonstrate a few useful implications for practice. For the first point, we have demonstrated that for these two networks, the detection accuracy performs very well, initially increasing very quickly and becoming better and better as more illness report data is gathered. In particular, the detection accuracy improves almost linearly with the number of contaminated nodes, though it will even continue to improve as cases accrue at the same locations. This means that in many outbreak scenarios, we will correctly identify the true source after only a very limited number of illnesses have presented, and as the number of illnesses increases, we identify the true source almost with certainty. Second, we have demonstrated that the detection performance is better, improving more quickly, for the *High Variance* network, which differs from the *Zero Variance* network only in its stochastic distribution of links out of each node. Since real food distribution network structures demonstrate great degrees of heterogeneity, not only in link distribution but in all parameters including flow distribution and variability in number of nodes per layer, our result provides both useful and positive information for practice. We might expect the accuracy to improve with increasing degree of heterogeneity in network structure, and thus with conditions observed if implemented in practice. Third, we have shown that it is possible to quantify the variability in detection performance, and that this information is useful for making specific statements bounding the accuracy of the algorithm's output. This reveals an important tradeoff for an investigator: wait for a certain number of illnesses to accrue until the source can be uniquely identified with very high accuracy, or act early to prevent further illnesses and implicate a greater, but still relatively small, and importantly, well defined number of top ranked candidates. Finally, we have seen that the algorithm's convergence process is independent of the prior probability assignment, even when that assignment reflects what we know to be actual outbreak generating behavior. This insight has an important practical implication, which is that the initial volume distribution over sources is not a necessary input to the traceback algorithm. In other words, the traceback methodology can be implemented with less information to achieve the same accuracy.

2.4. Conclusions

In this chapter, we formulated an original network-theoretical framework for identifying the location and initiation time of the source of a large-scale outbreak of foodborne disease. This framework takes into account the unique complexities that distinguish this problem from source identification in other network contexts. It assumes knowledge of the underlying distribution network and the set of observed illnesses at specific network locations and reported times. We developed an approximate inference algorithm for solving the source identification problem that exploits the temporal and structural dimensions of this problem.

We subjected the spatio-temporal traceback technique to initial studies to evaluate its performance. First, we developed a simulation-based evaluation framework that allows us to measure the success of the algorithm across a wide range of outbreak scenarios and network structures. We then applied this performance evaluation framework to two illustrative numerical examples based on stylized network structures. On first order, the results provide an indication of the accuracy and applicability of the method, and serve as an initial validation of intuitive properties. The application also yielded several important insights. As expected, detection performance is superior for networks with greater variance in degree distribution (i.e. the connectedness of nodes); this structural feature was designed into the algorithm's approach. Also unsurprisingly, variability in detection performance is low when accuracy is high. Moreover, it is possible to quantify the variability in detection performance. This has important practical implications, as it is useful for making specific statements that bound the accuracy of the algorithm's output. A more surprising result is that the algorithm's convergence process is independent of the initial conditions defining the outbreak contamination event, i.e., it does not depend to the choice of prior distribution over the sources. This result highlights the importance of network structure in detection performance. Furthermore, it means that the traceback methodology can achieve the same accuracy with less required input data. Taken together, these insights provide a first step in understanding how the accuracy of detection depends on the structure of a network and on the stochastic evolution of the disease trajectory.

While promising great potential to accurately identify the source of large-scale outbreaks of foodborne disease, the results reported in this chapter are based on stylized models and conclusions can only go so far. In the following chapter, we demonstrate that the performance exhibited here is robust when applied to network structures informed by real data. Furthermore, we show that the method can result in significant benefits in accuracy, specificity, and efficiency when compared with existing methods in outbreak identification.

Appendix 2.1.
Bayesian Network Traceback

In this appendix, we take a time-independent approach to identify the source based on the topology of the distribution network and the reported locations of contamination. The key idea of this framework is to view the problem from the perspective of a probabilistic graphical model (PGM), where each node is a random variable denoting whether the contamination has passed through that site (of food production, distribution, or retail). The PGM represents, through a set of conditional probability distributions, how the observation of a contamination at a given node increases the probability that the contamination has traveled through adjacent upstream and downstream nodes. The probabilistic model is used to identify the source of contamination as the node that maximizes the joint likelihood of the set of reported nodes. We start by introducing the process for generating the PGM from the distribution network *G*.

A1.1. Probabilistic contamination model

We introduce the directed probabilistic graphical model of contamination spreading, G', generated through a transformation to the graph $G\{V,E,N\}$. Each node $u \in V$ is modeled by a binary random variable, X_u, whose binary status, 1 or 0, represents whether the contamination has passed through node *u*. Edges $(u,v) \in E$ thus become direct probabilistic dependencies between X_u and X_v, representing how the presence of the contamination at v is linked to the probability of the contamination at u. We say that $X_u \in V(n)$ is a parent of $X_v \in V(n+1)$ if there is an edge $(u,v) \in E$, and refer to ω_v as the set of parents of X_v. We assume that X_v depends only on its parents, and therefore that G' is Markovian. It is important to note that even though each node takes on an "all or nothing" binary value, the model does not require that contaminated product arriving at a node cross-contaminates all product at that node; rather, proportions of contaminated product traveling from source to sink are propagated through the calculation of the probability of a contamination cascade, made using the joint probability distribution introduced below.

Defining the probability distributions is the key step in transforming G into G'. First, each node $s \in V(1)$ is assigned an unconditional prior probability distribution $P(X_s = x_s)$, representing the likelihood that s is the contamination source. These distributions are set according to the predefined prior probability distribution in (0), such that $P(X_s = 1) = p_s$ and $P(X_s = 0) = 1 - p_s$.

A conditional probability distribution is defined for each node *u* in stages $n > 1$, given the values of its parents, $P\left(X_v = x_v \middle| \{X_u = x_u\}_{X_u \in \omega_v}\right)$. We assume that the (conditional) probability the contamination has passed through v is equivalent to the proportion of $\langle f_v \rangle$, the time-average volume at v, sent from contaminated parents:

$$P\left(X_v = x_v \middle| \{X_u = x_u\}_{X_u \in \omega_v}\right) = \frac{\sum\limits_{X_u \in \omega_v} x_u \langle f_u \rangle f_{uv}}{\langle f_v \rangle}. \quad (7)$$

To calculate the time-average volume $\langle f_v \rangle$, we must first introduce *F*, a transition probability matrix of dimension $|V| \times |V|$ composed of the normalized flow proportions f_{uv}, such that $[F]_{u,v} = f_{uv}$. Elements along the diagonal of F correspond to self loops and are thus equal to 0, with the exception of the flows corresponding to the retailer nodes $w \in V(N)$, which are represented in F as absorbing states with probability 1. F is thus a proper right stochastic matrix, with each row summing to 1. We also define F_0, a matrix of dimension $|s| \times |s|$ with diagonal elements equal to the normalized initial volumes f_u at nodes $s \in V(1)$, and all other elements set to zero. Finally, the cumulative volume per unit time $\langle f_v \rangle$ at v is found as the sum down the v^{th} column of the result:

$$\langle f_v \rangle = \sum_{u \in V(1)} \left[(F_o)(F^n) \right]_{u,v} . \qquad (8)$$

After the initial and conditional probability distributions have been defined, the joint probability distribution over all the variables $X_v \in G'$ can be determined. Due to the Markov property, the joint probability distribution factorizes into a compact representation, computed using the formula:

$$P\left(\{ X_v = x_v \}_{X_v \in G'} \right) = \prod_{X_u \in G'} P\left(X_v = x_v \middle| \{ X_u = x_u \}_{X_u \in \omega_v} \right) . \qquad (9)$$

This joint distribution allows the evaluation of inference questions by marginalization, or summing out the unused variables. The transformation of a food distribution network G into G' is illustrated using a toy example in Figure 3.

A1.2. Probabilistic source identification problem

In the event of an outbreak of foodborne contamination and observations of illness $i \in O$, our aim is to use the joint probability distribution in (9) to identify the source of contamination, s^*. We assume the same contamination diffusion and source reporting processes as introduced in 2.1, with the following caveats. First, we ignore the contamination times and only look at the contaminated retailer nodes $o_i \in H$, with distributions X_{o_i}. Second, when $X_s = 1$, all nodes $u \in \Omega \setminus s$ by necessity take on the value 0, i.e. $\{ X_u = 0 \}_{u \in \Omega \setminus s}$. Thus, we aim to solve:

$$\hat{s} = \arg\max_{s \in \Omega} P\left(X_s = 1, \{ X_u = 0 \}_{u \in V(1) \setminus s}, \{ X_{o_i} = 1 \}_{o_i \in H} \right), \qquad (10)$$

which can be calculated from the joint probability distribution in (9) by marginalizing out the unused variables $\{ X_v \}_{v \in G \setminus H \cup \Omega}$ over possible values $\{0,1\}$, i.e.:

$$P\left(X_s = 1, \{ X_u = 0 \}_{u \in V(1) \setminus s}, \{ X_{o_i} = 1 \}_{o_i \in H} \right) = \sum_{\{ X_v \in \{0,1\} \}_{v \in G \setminus H \cup \Omega}} P\left(\{ X_v \}_{v \in G \setminus H \cup \Omega}, X_s = 1, \{ X_u = 0 \}_{u \in V(1) \setminus s}, \{ X_{o_i} = 1 \}_{o_i \in H} \right) .$$
$$(11)$$

The objective in (10) effectively models the bulk diffusion process, since the observation of contamination at a given node is taken to increase the conditional probability that the contamination has traveled through adjacent nodes. By tabulating over the probability of feasible transitions between nodes, it considers the collection of all possible cascades from source s to observations H.

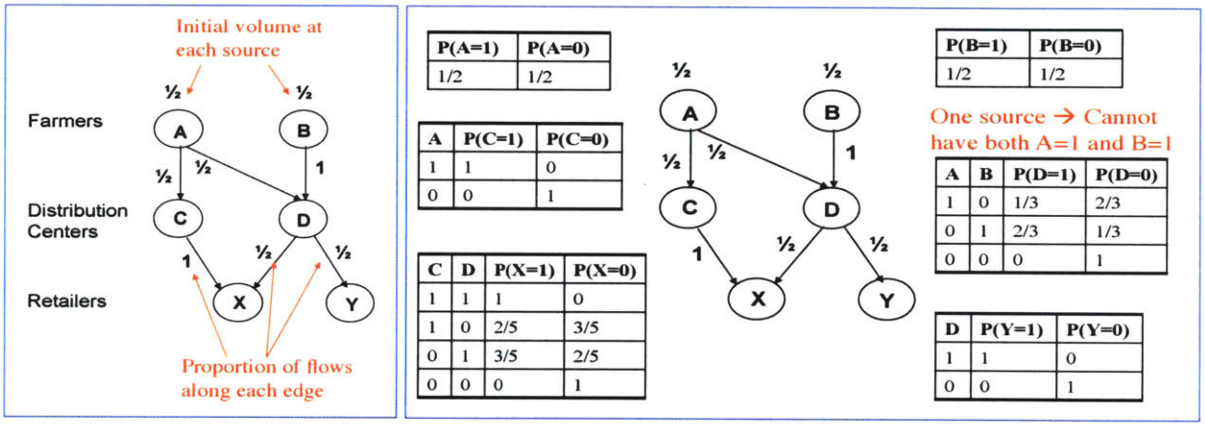
Finally, we can form a PMF over the possible sources,

$$P\left(s^* = s \middle| \{ o_i \}_{o_i \in H} \right) = \left\{ \; P\left(X_s, \{ X_u \}_{u \in V(1) \setminus s} \middle| \{ X_{o_i} = 1 \}_{o_i \in H} \right) \text{ for } s \in \Omega \; \right\} \qquad (12)$$

by normalizing the joint probabilities in (10) over the set $s \in \Omega$.

Table 2. Bayesian Network Source Detection Algorithm
Inputs:
$G\{V,E,N\}$, food distribution network

```
        $P(s^* = s)$, prior distribution
        $O$, observation set as of time $t_W$
        $\tau$, contamination time uncertainty
For $u \in V(1)$:
        Determine if $u$ reaches all observed contamination nodes $o_i \in H$, then return $s \in \Omega$
Define graphical model $G'$:
        For $s \in \Omega$:
            Set $P(X_s = 1) = p_s$ and $P(X_s = 0) = 1 - p_s$
        For $v \in G \setminus H \cup \Omega$:
            Determine conditional probability distributions $P\left(X_v = x_v \middle| \{X_u = x_u\}_{X_u \in \omega_v}\right)$ using Equation (7)
Return
        $\hat{s} = \arg\max_{s \in \Omega} P\left(X_s = 1, \{X_u = 0\}_{u \in V(1) \setminus s}, \{X_{o_i} = 1\}_{o_i \in H}\right)$, computed using (11).
```



Initial volume at each source
½
½
Farmers
A
B
½
½
1
Distribution Centers
C
D
1
½
½
Retailers
X
Y
Proportion of flows along each edge
½
½
A
B
½
½
1
C
D
1
½
½
X
Y
One source → Cannot have both A=1 and B=1

P(A=1)	P(A=0)
1/2	1/2

A	P(C=1)	P(C=0)
1	1	0
0	0	1

C	D	P(X=1)	P(X=0)
1	1	1	0
1	0	2/5	3/5
0	1	3/5	2/5
0	0	0	1

P(B=1)	P(B=0)
1/2	1/2

A	B	P(D=1)	P(D=0)
1	0	1/3	2/3
0	1	2/3	1/3
0	0	0	1

D	P(Y=1)	P(Y=0)
1	1	0
0	0	1

Figure 3. Transformation of a food distribution network into a probabilistic model. Consider the simple food distribution network model made up of 2 Farmers, 2 Distribution Centers, and 2 Retailers. The conditional probability tables are then computed using equation (7). The joint probability distribution for this model can be written in factored form as:

$$P(A,B,C,D,X,Y) = P(A)P(B|A)P(C|A)P(D|A,B)P(X|C,D)P(Y|D).$$

Appendix 2.2: References documenting residence times and network structure

Data source / Paper	**Times / stages of the supply chain**	**Relevant variables included** (Time spent, detail on supply structure and flows)
Compiled research on the head lettuce supply chain, from farm to fork, compiled by BT Safety LLC 2005	This data, collected on food in 26 product categories from multiple industry partners, consists of shipping records and compiled results from expert elicitation. The expert elicitation is a compendium of interviews, research, and industry knowledge. The shipping records document movement from origin to distribution center and distribution center to retail warehouse for one large foodservice company, as well as a set of retail-only packing/shipping records that make up a subset of one large company's production. Because this data is based on actual records of product inflow and outflow for a compendium of companies, taken together, it unveils distribution patterns. **Relevant parameters:** · Industry averages: o Time spent in field, on average o Volume in field, on average **Match these to distributions from Green Giant data o Harvest size, on average o Shelf life o Average residence times at point of harvest, cold storage, distribution center storage, at retail, and a good model for home storage (see percentages; they used lognormal(avg = 1.3 days)) o Method for computing transportation times, based on long-haul trucking service regulations · For one specific retailer: o Specific volumes from origin nodes – distributors – customer zipcode o Degree distributions from origin zipcode to distributors	
Pinior, Beate, Uwe Platz, Ulrike Ahrens, Brigitte Petersen, Franz Conraths, and Thomas Selhorst. "The German milky way: trade structure of the milk industry and possible consequences of a food crisis." *Journal on Chain and Network Science* 12, no. 1 (2012): 25-39.	Study analyzing the structure of the aggregated trade network of the milk supply in Germany and the public health risk of contamination.	**Relevant parameters:** · The number of nodes in the supply chain (producers, dairies, consolidation companies, consumers/retailers) · The degree distributions (distributions of links between actors in the supply chain), including max, min, and average

(Beate Pinior's PhD thesis provides further information on this supply chain and her risk analyses)		degree · Documents data sources for trade volumes but does not plot actual distributions
Beni, Leila Hashemi, Sébastien Villeneuve, Denyse I. LeBlanc, and Pascal Delaquis. "A GIS-based Approach in Support of an Assessment of Food Safety Risks." *Transactions in GIS* 15, no. s1 (2011): 95-108.	Study analyzing the structure of the aggregated trade network of bagged lettuce in Canada and the public health risk of contamination. · Collected the data to model the aggregated geo-coded, time-sensitive distribution chain for ready-to-eat bagged salad in Canada · Created a simulation to model a contamination spreading through the food supply.	**Data collected by authors (but not provided in the paper)**: · Number of establishments involved in the supply chain · Links between establishments · Flows (quantities) of product moving between establishments · Residence and transportation times
LeBlanc, Denyse I., Sébastien Villeneuve, Leila Hashemi Beni, Ainsley Otten, Aamir Fazil, Robin McKellar, and Pascal Delaquis. "A national produce supply chain database for food safety risk analysis." *Journal of Food Engineering* 147 (2015): 24-38.	Extension of the above study; adds in temperature and microbiological growth.	**Relevant parameters:** · Provides probability distributions for residence and transportation times of lettuce through the supply chain · Also includes nodes per stage in the network
McKellar, Robin C., Denyse I. LeBlanc, Fernando Pérez Rodríguez, and Pascal Delaquis. "Comparative simulation of Escherichia coli O157: H7 behaviour in packaged fresh-cut lettuce distributed in a typical Canadian supply chain in the summer and winter." *Food Control* 35, no. 1 (2014): 192-199.	· Study used to inform LeBlanc's time distributions. · Measured residence times of lettuce in winter and summer months in a retail supply chain from a processing facility to retail storage	**Relevant parameters**: · Storage and transportation at Producer, Distribution Center, and Retail
Danyluk, Michelle D., and Donald W. Schaffner. "Quantitative assessment of the microbial risk of leafy greens from farm to consumption: preliminary framework, data, and risk estimates." *Journal of Food Protection* 74, no. 5 (2011): 700-708.	· Monte Carlo simulation to model microbiological risks associated with the growth/decay of specific pathogens in fresh-cut lettuce/leafy greens throughout supply chain · A few distributions from data, a few they made up	**Relevant parameters:** Time in field after contamination Retail storage time Home storage
Pouillot, Régis, Meryl B. Lubran, Sheryl C. Cates, and Sherri	· Determined probability distributions of time food spends	**Relevant parameters**: Home storage time distributions for

Dennis. "Estimating parametric distributions of storage time and temperature of ready-to-eat foods for US households." *Journal of Food Protection®* 73, no. 2 (2010): 312-321.	in home storage by fitting consumer survey data to survival distributions · Products: Many ready-to-eat products including bagged salad and soft cheese	bagged lettuce, soft cheeses
Laguerre, Or, H. M. Hoang, and D. Flick. "Experimental investigation and modelling in the food cold chain: Thermal and quality evolution." *Trends in Food Science & Technology* 29, no. 2 (2013): 87-97.	Survey article of deterministic and simulation methods for modeling relationship between time-temperature and pathogen growth throughout the cold chain (for various foods)	· **Relevant parameters:** Parametric distributions for times spent in transport, storage, hub, cold room, display cabinet, transport by consumer, and at home (domestic refrigerator)
Dallaire, R., D. I. LeBlanc, C. C. Tranchant, L. Vasseur, P. Delaquis, and C. Beaulieu. "Monitoring the microbial populations and temperatures of fresh broccoli from harvest to retail display." *Journal of Food Protection* 69, no. 5 (2006): 1118-1125.	· Product: Broccoli · Small case study tracking broccoli through a supply chain involving 2 growers, 1 wholesaler, 4 retailers in same geographical area; measuring microbial population, temperature, and time spent	· **Relevant parameters**: Producer storage, wholesaler storage, retailer storage. In particular: · Storage on farm after harvest: 1.8 days avg., max 4 days · Storage at wholesaler: less than 24 hours · Storage at retailer before display: Min: few hours; max: 7 days; avg: 1-3 days for store 1, <1 day for store 2

Chapter 3: Robustness and Benefits

In Chapter 2, we presented a network-theoretical framework for identifying the source of large-scale outbreaks of foodborne disease. Through application to numerical examples, we demonstrated an initial indication of the methodology's accuracy and a first step towards understanding how the accuracy of detection depends on network structure. Our ultimate objective is to evaluate the utility of these methods in real-world scenarios, and in comparison to existing methods in outbreak investigation. In this chapter, we develop and apply the traceback methodology to (i) a set of network models representing key structural and spatial properties of food distribution networks in the US and (ii) a framework for quantifying benefits in comparison to existing approaches to foodborne disease outbreak source detection: to a practical baseline meant to demonstrate current methods applied in practice during outbreak investigations, and to a best-in-class method presented in the literature.

The contributions of this chapter fall into four categories:

- We identify key structural and spatial features characterizing food distribution networks and develop a generalizable modeling framework that uses available data to represent these features.
- We show the traceback methodology robustly identifies and localizes the outbreak source when applied to realistic network structures; these results serve as a first step in validating its application in practice.
- We extract generalizable conclusions regarding the dependence of traceback accuracy on network structural features.
- We show that the method can result in significant benefits in accuracy and efficiency when compared with existing approaches in foodborne disease outbreak source detection.

The chapter is organized as follows. In Section 3.1, we develop the network modeling framework and implement it to model the distribution of two specific commodities in the US: tomatoes and lettuce. In Section 3.2, we evaluate the performance of the traceback methodology applied to these two realistic structures. In Section 3.3, we quantify the benefits. Section 3.4 concludes.

3.1. Modeling Real Distribution Networks

Our objective is to validate the traceback framework through its successful application to real-world distribution networks and outbreak scenarios. In Chapter 5, we propose a method for developing a live, accurate distribution network model for specific foods from publicly available data sources. At the time of preparing this dissertation, granular supply chain data representing distribution networks for various commodities was not available. In the absence of exact data, our aim was to design a set of network models representing at a high level the realistic spatial and structural features of distribution networks for specific foods in the US. Through a review of the literature on food distribution networks, we identified four key features characterizing real distribution networks: consolidation of the food industry in specific clusters of production, spatial scale and long-ranging distribution links, heterogeneity of volume and degree distributions, and structural relationships between stages. We represent each of these features in the model, using statistical data to inform parameter choices wherever possible. Structural assumptions and probability distributions were used to fill in where this data was not available.

3.1.1. Representing Key Structural Features

A review of distribution structure identified the following four key features characterizing food distribution networks. In this section we introduce the features and describe our modeling approach for representing them in the network models. These features were developed based on data from government published sources (USDA NASS 2012; USDA NASS 2016; USDA ERS 2016; AMS 2016), academic literature (as cited below), and through information elicited in interviews with industry experts (Appendix 2.2).

Consolidation of the industrial agricultural system in production and processing; spatial scale and long-ranging distribution

In past decades, the food system in the US has undergone dramatic consolidation, or the organization of production into fewer but larger farms or plants. Consolidation is occurring in both production or farming practices and processing. While population and consumption levels have expanded enormously, the total number of US farms has declined from 5 million farms in 1950 to about 2.2 million in 2010. Processing plant sizes have increased sharply, facilitated by improving processing plant technologies and the emergence of mega-corporations and new "scale economies." An extreme example can be seen in the poultry industry in the US, where 95% of chickens produced for meat are grown under production contracts with fewer than 40 companies. (Nesheim et al. 2015; MacDonald et al. 2013).

Consolidation and concentration in the food supply chain have important implications for modeling distribution structure, and in particular, on the spatial distribution of production and processing nodes. As the industry consolidates, the location of supply chain actors will cease to be distributed throughout the country as the sites of production and processing aggregate at specific centralized locations. The result is a supply chain where nodes are organized in fewer, larger "clusters" of production. The distance between supply chain nodes plays a lesser role in the choice of trade partners.

To reflect these observations in the spatial structure of nodes and connections in our models, we adopt the following approach.

- The major geographical clusters accounting for the top 80 – 95% of production and processing volume are represented (USDA NASS 2012; 2016; USDA AMS 2016); these clusters determine the location of production and processing nodes.
- Consumption is represented in proportion with population (USDA ERS 2016); this determines the location of warehouse and retail nodes.
- The incidence of trade connections reflect these assumptions, with proportionally more out-degree links for clusters producing greater volume and more in-degree links in for regional areas of greater population.
- In matching supply and demand, we implement a bias to connect nodes locally. However because the clusters of production are not evenly spatially distributed throughout the country, regional supply will often not satisfy regional demand. As a result, the majority of supply and demand is matched according to volume capacity alone, without considering the distance between node locations.

Heterogeneity in structure

As suggested in previous chapters, food distribution networks are characterized by heterogeneity in (i) the distribution of the number of links leaving each node across all nodes in a stage, or the out-degree distribution, (ii) the initial volume distribution across producing nodes the first stage, and (iii) the distribution of flow volumes across the links leaving a single node, or the flow distribution. This behavior has been observed in network studies documenting supply chain structure (Manitz et al. 2014; Pinior et al.

2012, Friedrich 2010), and moreover characteristic of complex networks in general (Brockmann and Helbing 2013; Grady et al. 2011).

To represent heterogeneity in out-degree distributions, we generate links according to the geometric probability distribution, $X_{\text{out}} \sim \text{Geom}\left(\frac{1}{\mu}\right)$, where μ is the average out-degree. Where available, μ is fit to distributional data. To represent heterogeneity in the initial volume distribution $F_o(s)$ across the v_1 producing nodes in stage 1, we sample volumes from the actual farm size distribution data published by the USDA (USDA 2012). To represent the flow distribution across links leaving a node, we sample volume proportions according to the geometric distribution $F(f) \sim \text{Geom}(1/\lambda_F l_m)$, where l_m represents the number of links leaving node *m* and scale parameter $\lambda_F = 5$.

Supply Relationships Across Stages

The structure of the supply chain is characterized by a series of relationships with respect to a scale-up or a scale-down in the number of nodes from one stage to the next (Conrad et al. 2012; Pinior et al. 2013; LeBlanc et al. 2015). While the degree of scale-up or down will depend on the commodity, in general the following relationships are observed: (i) the number of nodes from grower to processor stage is a scale-down relationship; (ii) the number of nodes from processor to warehouse stage is a dramatic scale-up relationship; and (iii) the number of nodes from warehouse to retailer is a dramatic scale-up relationship.

These relationships are useful in modeling the number of nodes v_n in each stage n. The number of farm or producer nodes for various commodities are available from USDA Census of Agriculture (USDA NASS 2012; 2016) and the number of retailer nodes are informed by the USDA Economic Research Service (USDA ERS, however the number of interior stage nodes (i.e. processor and warehouse) are not readily available. The scale-up and scale-down relationships thus help to set parameter values that reflect general structural characteristics.

3.1.2. Model Implementation

We have developed models based on the structural and spatial features highlighted above to the distribution of two commodities: Tomatoes and Lettuce. In this section we overview the modeling choices and parameter values implemented.

Foods Chosen

To simplify the structure of the distribution network and focus the analysis of traceback accuracy results, we model two single-ingredient fresh produce commodities: Tomatoes and Lettuce. Both of these distribution structures can be represented by the four stages of supply introduced in Chapter 2: Farms, Processors, Warehouses, and Retailers. The primary reason for choosing these commodities was structural. In particular, lettuce was chosen because it exhibits one of the most consolidated supply structures, with 95% of production being concentrated within 2 growing regions: central California and Yuma, Arizona (USDA 2015; AMS 2015). Extreme clustering in production makes it more difficult to distinguish between feasible source nodes in the event of an outbreak, since distance is no longer a differentiating factor. We thus see this network as representing a "lower bound" on the accuracy achievable with the traceback methodology. If we are able to achieve accurate results with this network, it would suggest that accuracy will be higher with other commodity types. Tomatoes (in summer season) were chosen because they represent an average distribution of production clustering throughout the country.

In addition to structural considerations, these two commodities have a history of carrying foodborne disease (*E. coli*, *salmonella*, and *listeria*) that has escalated in recent years, and are thus ripe for studies to improve food safety and faster detection of outbreak sources. Because of this history of outbreaks, understanding the structure of these two commodities has been high priority for investigators, from whom we have been able to gather considerable background information from both directly and from past outbreak reports.

Another consideration is that of the generalizability of the distribution models based on fresh produce-specific supply chains. Fresh produce was chosen above other commodities due to the purity of product. While extensions to more complex, processed products can be imagined, the scale up would introduce considerable complexity that may not be accounted for in our modeling framework. As one example, some processed products have many month long shelf lives and may even be stored somewhere in the middle of the supply chain, which completely invalidates the use of temporal parameters to track the product's trajectory. Still, because produce is easier to model it is a good starting point for more complex modeling frameworks. Furthermore, fresh produce on its own is a significant and increasing contributor to the disease burden from foodborne disease, with recent estimates citing produce as the cause of 13.5% of illnesses, 13.6% of hospitalizations, and 10.4% of deaths attributed to foodborne disease (Batz et al. 2012).

Spatial and Population Assumptions

The geography, population distribution, and production clusters modeled in each network are pictured in Figure 3.1. The networks exist on a geographical area extending 2500 x 1500 miles, roughly resembling the structure of the US. This area is divided into three regions with grid boundaries based roughly on US time zones: Pacific and Mountain as one region, Central as a second, and Eastern as a third. Population fractions in each region are set to reflect the population of the states comprising each time zone, with each region divided into sub-regions as depicted in the figure. Warehouse and Retailer nodes are allocated across each region and then across each sub-region according to population, while ensuring that there is at least 1 Warehouse node per sub-region.

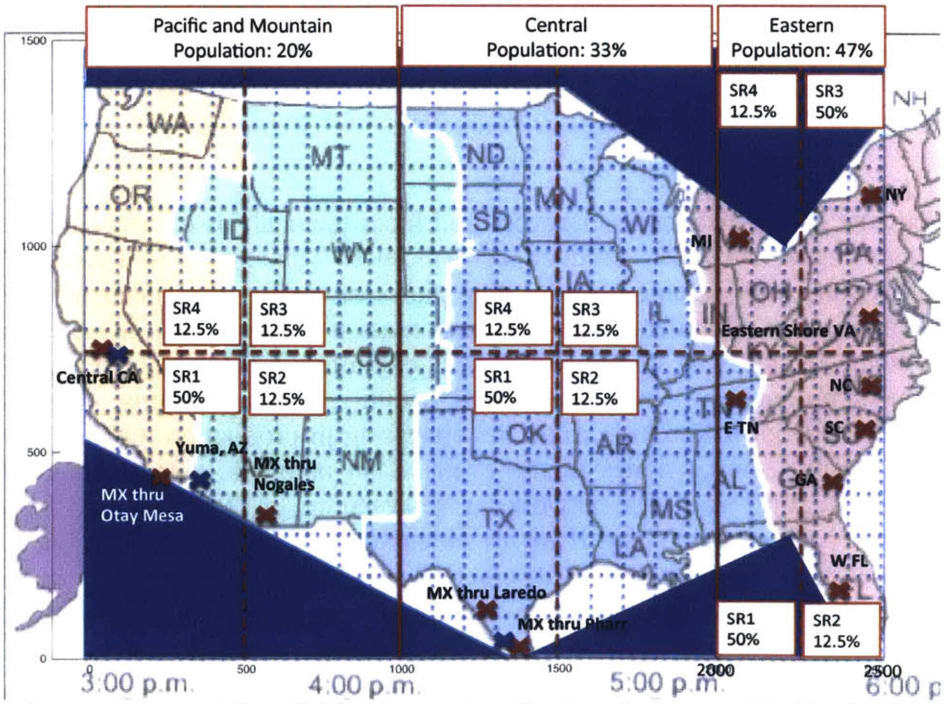


Figure 3.1. Geography, population distribution, and production clusters modeled in the Tomato and Lettuce distribution networks. Regional (sub-regional) populations as a fraction of US (regional)

totals. Red and blue X's in Figure 3.1 mark the location of tomato and lettuce growing clusters, respectively.

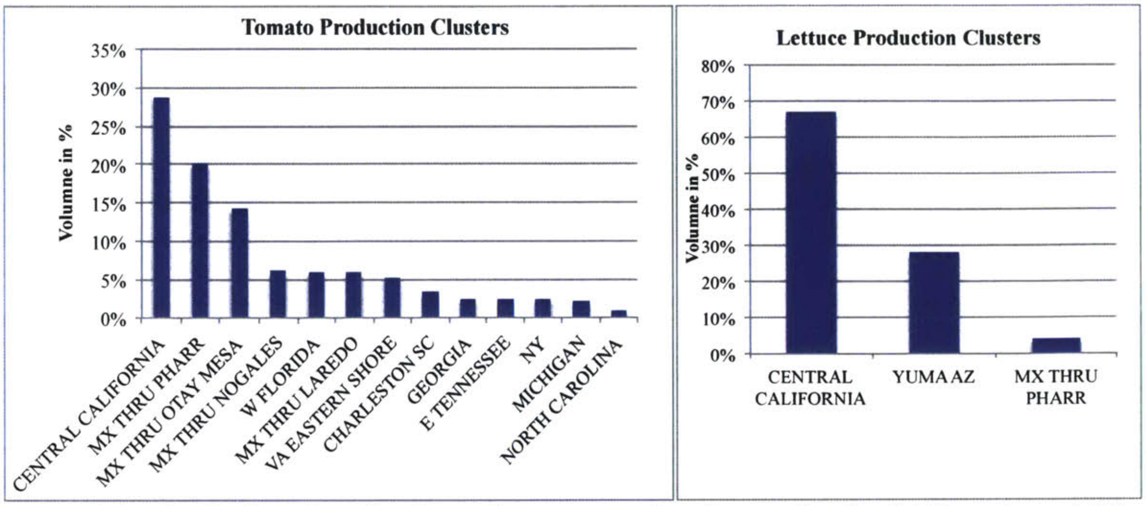
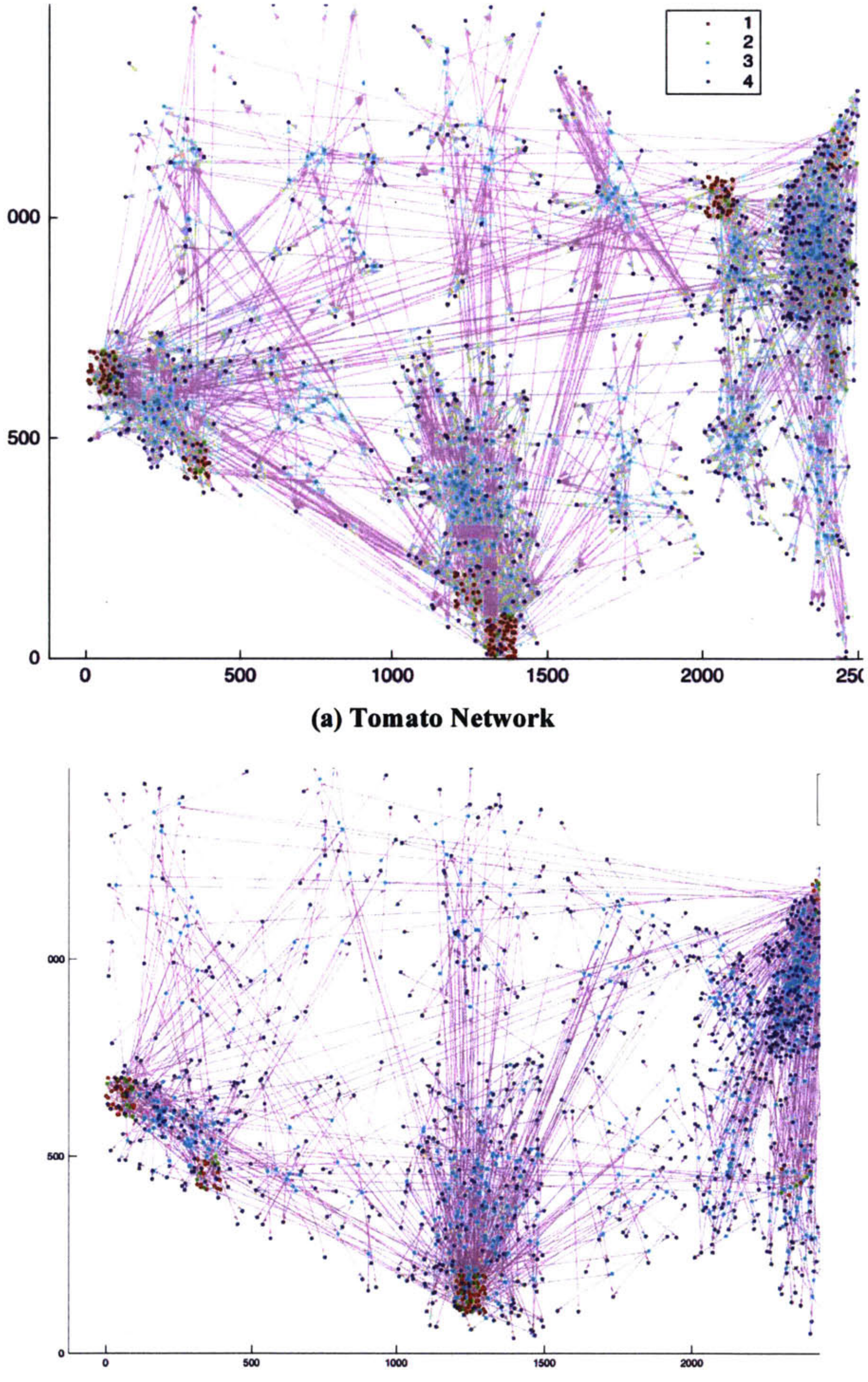


Figure 3.2. Distribution of production volume across the clusters for the **(a)** Tomato and **(b)** Lettuce networks.

For the reasons discussed above, we model farming and processing of each commodity as being concentrated in a variable number of clusters of production in the US and Mexico (via border crossings). Each cluster is modeled as a 100 x 100 grid box in the map. We assume that production of raw commodities is minimal and occurs at processing centers very close to the site of production. Thus, trade between Farm and Processor nodes occurs only within clusters. We capture ~80% of the production volume of tomatoes by representing the top 13 clusters of production, including 4 border crossings from Mexico (USDA 2015; AMS 2015). For lettuce, we represent the California and Arizona clusters that together account for ~95% majority concentration of production, adding a third cluster at a Mexico border crossing to capture ~99% of production. The red and blue X's in Figure 3.1 mark the location of tomato and lettuce growing clusters, respectively. Figure 3.2 reports the distribution of production volume across the clusters for each network. The network structural parameter values realized from implementing the modeling approach with these clustering and population assumptions are reported in Table 3.1. The resulting networks are visualized in Figure 3.3 a and b.

	Tomato Network	**Lettuce Network**
Clusters of production		
	13	3
Nodes v_n in stage *n*		
Farms, v_1	1500	500
Processors, v_2	250	100
Warehouses, v_3	1500	1500
Retailers, v_4	2500	2500
Average out-degree μ_n for stage [*n*, *n+1*]		
Farm - Processor	2	2
Processor - Warehouse	12	30
Warehouse - Retailer	5	5

Table 3.1. Model parameters for the Tomato and Lettuce networks.



(a) Tomato Network

(b) Lettuce Network

Figure 3.3. Visualization of the **(a)** Tomato and **(b)** Lettuce network models. Farms are represented by red nodes, Processors by green nodes, Warehouses by blue nodes, and Retailers by purple nodes.

3.2. Traceback Accuracy for Regional Network Models

In this section we evaluate the performance of the traceback methodology applied to the Tomato and Lettuce networks. We implement the same performance evaluation framework used in Chapter 2, generating 100 outbreak simulations according to the baseline scenario specifications. Results are assessed according to the Traceback Accuracy metric. Since the networks are represented on the spatial dimension, we add an additional metric to measure the ability of the methodology to geographically identify the source, the Distance From True Source. As its name implies, this metric measures the geographical distance between the algorithm's top ranked source $\hat{s}$ and the true source s^*. We report source identification and localization performance as a function of the number of illnesses; this is the most granular interval of progression, since multiple illnesses can occur on the same day and at the same node.

3.2.1. Results: Source Identification and Localization

Source Localization

Figure 3.3a plots the Distance from the True Source as a function of the number of illnesses reported. The variability in the simulation results is quantified by the PMFs pictured in Figure 3.3 b and c, which plot the Distance from the True Source as a random variable D for various intervals in number of illnesses reported. Localization performance is precise and efficient. After only 20 reports of illness, the true source is identified within ~10 miles for the Tomato network and ~2 miles for the Lettuce network, on average; a focused area given the entire country is being considered. While the variable in results can range well beyond the averages, the method is able to localize the source within 30 and 50 miles, for Tomatoes and Lettuce respectively, in 100% of simulations after 20 illness reports. Because each cluster is modeled as a 100 x 100 grid box in the map, this means that the source cluster is accurately identified in all simulations.

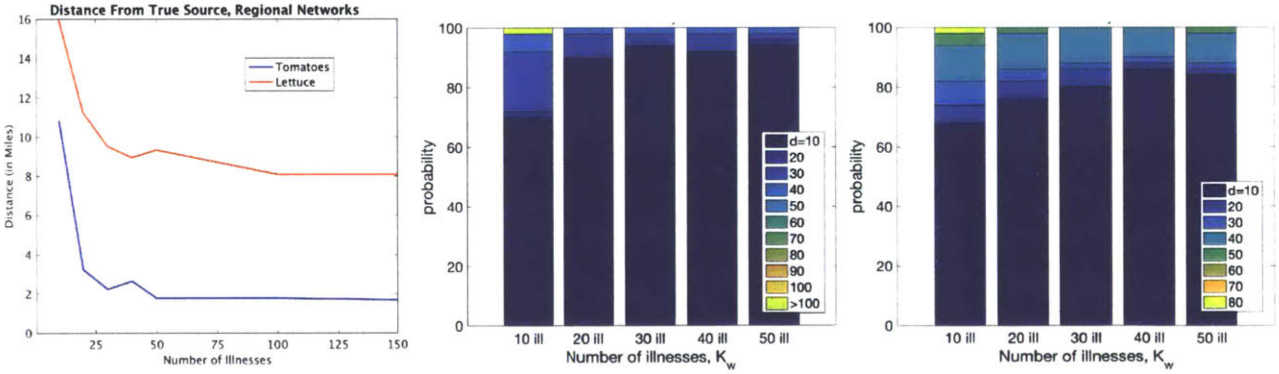


Figure 3.3. (a) Distance from True Source for Tomato and Lettuce networks, as a function of the number of illnesses **(b)** Probability mass functions for Rank of True Source R, at intervals in number of illnesses K_ω = 10, 20, 30, 40, 50 for the Tomato network and **(c)** Lettuce network.

Source Identification

Figure 3.4 a and b plots Traceback Accuracy (TA) and the estimation error of the start time $\hat{t}_s$ for the Tomato and Lettuce networks as a function of the number of illnesses. Traceback performance is both accurate and efficient, with TA converging to close to peak value after only 25 illnesses for both networks. (The variability of these results is quantified in Chapter 3, where we demonstrate detection rates over 80% accuracy after only 25 illnesses have been reported). The start time estimate improves more linearly with the number of reported illnesses, in both cases identifying the source within <3 days of the true start time; that is, within ± 1.5 days of the actual contaminated batch.

The rapid convergence to a peak accuracy value 10 – 20 % below 100% accuracy is attributable the densely connected clusters. Essentially what is happening is that the algorithm is able to rapidly narrow in on the correct cluster of distribution, as shown above, but then has greater difficulty in distinguishing between nodes within the cluster. The combination of the scale-down relationship between the grower and processor stages and the high density of connections within each production cluster means that there is a "bottleneck" structure, with multiple grower nodes connected to the same processor nodes. More specifically, the set of retailer node decedents of the true source node will be identical to the retailer node descendants for another source node within the same geographical cluster. This means that the problem is only reducible to a specific number of sources as dictated by the regionally dependent connectivity patterns of a given network structure. The start time, on the other hand, improves with each new piece of evidence (i.e. illness report at a given time) as dictated by the optimization function generating the estimate.

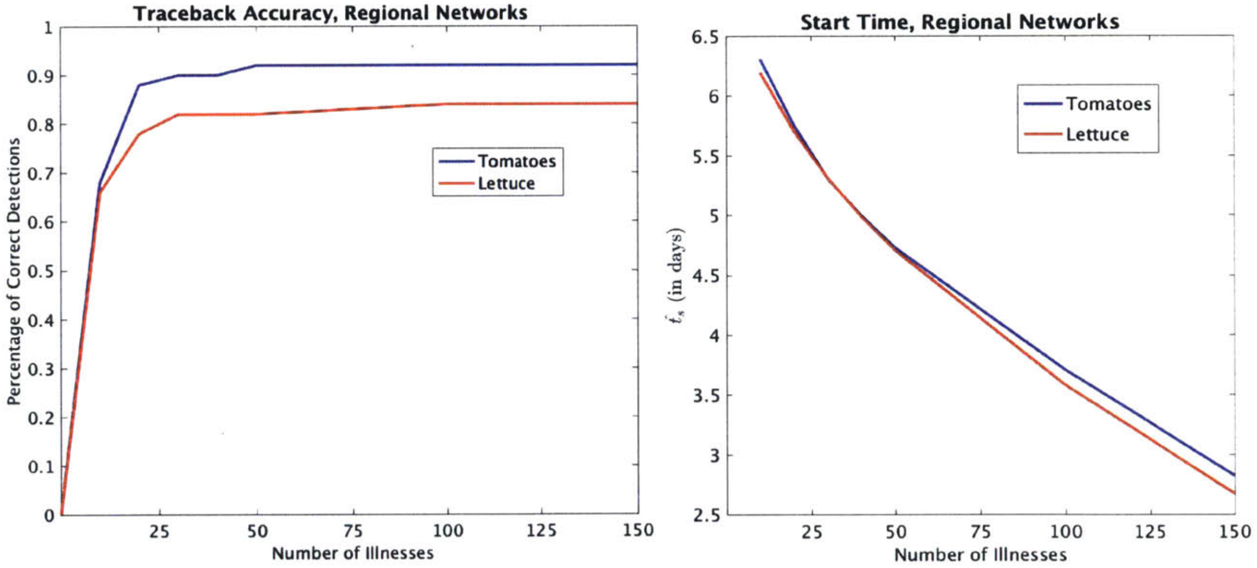


Figure 3.4: (a) Traceback Accuracy (TA) and **(b)** Estimation Error of the Start Time $\hat{t}_s$ as a function of the number of illnesses K_ω, for the Tomato and Lettuce networks.

Network Structure and Traceback Accuracy

We now comment on differences between the two networks and implications for understanding the role of network structure on traceback accuracy. As expected, traceback performance is better for the Tomato network, converging to 92% correct detections vs. 83% for the Lettuce network. The difference in accuracy can be attributed to some combination of the greater (i) density and (ii) clustering of the lettuce network. With regard to (i), we observe from the structural parameters reported in Table 3.1 that the Lettuce network is more densely connected, demonstrating a significantly higher average out-degree between the Processor and Warehouse stages. We expect this factor to also lessen the traceback performance of the Lettuce network in comparison with the Tomato network, since a greater link density means there are more possible network pathways to consider, increasing the uncertainty in the problem. With regard to (ii), we have discussed how the Lettuce network represents an extreme in the clustering of production for which two geographically close regions fulfill ~95% of US consumption. Due to the high degree of clustering, the time taken in distribution to reach any given contaminated node will be similar from all feasible sources, making the temporal signal sent by the true source less distinguishable. As a result, the start time likelihood objective will contribute less in the way of prediction to the source localization problem. At the same time, the similarity in distribution times will mean that an accurate start time estimate can be achieved, as exhibited in Figure 3.4b.

3.2.2 Practical Implications

In this section, we have demonstrated high accuracy in identifying and localizing the outbreak source when the traceback methodology is applied to geographically realistic network structures. The many network structural features influencing traceback performance can interact in complex ways, meaning that caution must be taken when deriving general conclusions regarding the accuracy of source traceback for foodborne disease outbreaks. Furthermore, it is important to stress that while the network models developed here are representative of key structural and spatial features and are based on current production and distribution data, they are still stylized, high-level models that are neither perfectly representative or complete. A much more complex trade network would result from the consideration of all supply chain actors, big and small, characterizing all commodity flows as well as external trade relations with different producers in the industry. Furthermore, our conclusions are derived from a simulation-based evaluation framework. The next step towards validation of the methodology will be to demonstrate its ability to correctly localize the origin of outbreaks in historical outbreaks. Live use of these techniques may demonstrate features of the real problem inadvertently omitted from the modeling.

Still, the results presented suggest a few important insights regarding the dependence of traceback accuracy on network structure, and the robustness of results for food distribution networks in general. First, we have seen that initial observations regarding the relationship between network structural features and traceback accuracy in 2.3 are robust when applied to these realistic networks; in particular, performance improves for networks characterized by a greater degree of heterogeneity, in this case in the geographical distribution of nodes. What this means in practice, if we generalize from these conclusions, is that we can expect higher traceback accuracy for foods that are produced across a *greater number* of *spatially distributed* locations. We have also seen that performance is worse where the density of connections is greater. This means Traceback Accuracy prefers industries where there are more supply chain actors who work with a smaller number of other entities. Still, as we have seen in Section 2.3, differences i.e. heterogeneity in connectivity is needed to best differentiate between sources, and thus a distribution of connectivity where some nodes are more highly connected than others will improve accuracy. Additionally, as discussed in Section 3.1, lettuce is one of the most aggregated products in the country, with 95% of our lettuce supply grown in only 2 small regional areas. That traceback performance performs well for this network means it has succeeded for a structure representing reasonable realistic bounds, i.e. "worst case" boundary values. This observation suggests that we can expect location performance at least as accurate for other fresh-produce food items. In the next section, we provide further support for the traceback methodology presented in this thesis by quantifying benefits in comparison to existing approaches.

3.3. Quantifying Benefits

We now quantify the benefits resulting from the network-theoretic framework and Spatio-Temporal Traceback (STT) algorithm developed in Chapter 2 and evaluated in application to realistic network structures in the previous section. We compare STT to existing approaches in foodborne disease outbreak source detection: a practical heuristic meant to model current methods applied in practice, and a best-in-class method presented in the literature (Brockmann and Helbing 2013). Differences in the Traceback Accuracy achievable with each method are measured to quantify benefits. If STT can significantly outperform the heuristic, it would suggest that the network-theoretical approach to source identification developed in this Chapter can contribute substantially to the traceback investigation process as a tool for tactical decision-making. The case for implementing our method is strengthened if it can also outperform existing theoretical approaches in the scientific literature.

3.3.1. Existing Approaches to Traceback

Modeling Current Methods in Investigation

The heuristic was developed from discussions with FDA investigators. It models the process applied to identify contamination sources during outbreak events, in the absence of a network-theoretic method like the methodology presented in this thesis. Specifically, investigators implement a process of "triangulation," or tracing back the distribution paths of products from several locations to determine if there is a common point of convergence in the supply chain. An example of a product trace diagram depicting exposure pathways is illustrated in Figure 1.2. As described in the introduction, because the set of all possible supply chain pathways leading to each chosen location must be traced independently along the supply chain without a structural network model to guide this investigation, the process is time and resource intensive. According to an estimate provided by investigators with the Minnesota Department of Public Health, 8 to 24 person hours are required to collect paperwork and create a product trace diagram for 1 to 2 contamination cases (Smith 2015). This information provides the basis for an assumption regarding the time necessary to perform product tracing for individual cases in the following heuristic. Of course, many potential differences with actual decisions might arise from practical factors that are not included in this model. For example, investigators will aim to identify a minimal convergence set by choosing cases that are part of distinct sub-clusters of contamination emerging at different restaurants or retailers, or choosing locations that are geographically distant from one another, e.g. a case in California, Texas, and Massachusetts. The model presented here chooses cases at random, without incorporating any information that might help to reduce the overlapping set. Discussions with outbreak investigators are required to design iterative improvements to this heuristic and its implementation.

FDA Heuristic

The FDA Heuristic is defined as follows. On each outbreak day t_ω, the decision-maker chooses at random 1 report of illness i linked to a unique node o_i from the set of reported illnesses K_ω. The choice o_i is added to the set of chosen contamination report nodes which we call the *triangulation contamination node set* H_T. For each node $o_i \in H_T$, the possible exposure pathways are traced back through the supply chain. The set of all common points of convergence is identified as the set of nodes in stage $n = 1$ that have at least one network pathway to the nodes $o_i \in H_T$; we call this the *feasible source set by triangulation* Ω_T. This process is continued as the outbreak progresses, so that by day $t_\omega = \tau$, there are $|H_T| = \tau$ nodes in the triangulation contamination set, where $|H_T|$ is equal to the size of the set H_T. The Traceback Accuracy on any given day t_ω is then determined as $1/|\Omega_T|$, where $|\Omega_T|$ is equal to the size of the set Ω_T on t_s.

Exploiting Network Structure

We also introduce a baseline meant to demonstrate the advantage of access to the network structural data while at the same time accentuating the value of the probability models incorporated into STT to differentiate between outbreak sources.

Network Baseline

The network baseline assumes that the underlying distribution network $G(V,E,N)$ is available and that the information from all reports of illness K_ω can be utilized to identify the full feasible source set Ω. Again, no mechanism is applied to use network structure to distinguish between sources $s \in \Omega$, and Traceback Accuracy is determined as $1/|\Omega|$. A possible but unimplemented extension of this baseline would involve differentiating between feasible sources by weighting them according to a prior probability distribution.

Existing Network-Theoretical Traceback Approach

Finally, we compare the method to the only known existing approach applying network-theoretic methods to identify the source of a foodborne disease outbreak, the "Effective Distance" method presented by Brockmann and Helbing in a 2013 paper. An overview of this method is provided in Section 1.3.2; readers should refer to the paper for further details. In the following, we refer to this method as B&H.

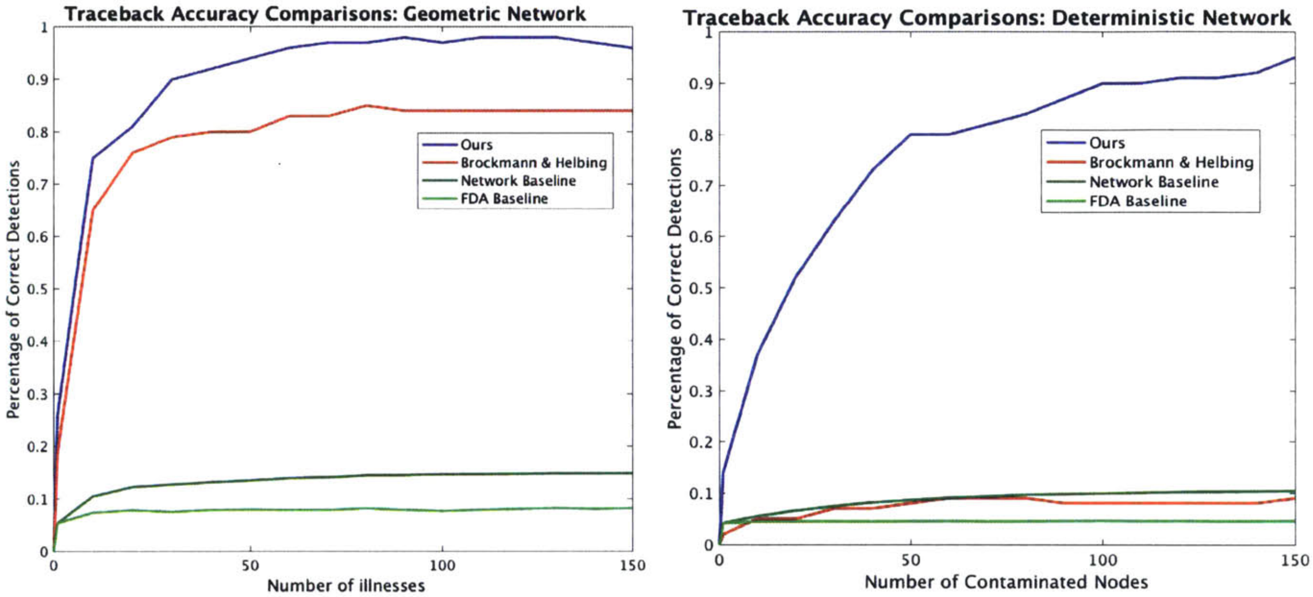
3.3.2. Comparisons

Numerical examples

First, we compare the detection performance obtained by applying each of the four traceback methods to the *Geometric-4* network and the performance evaluation framework introduced in Section 3.2.1. Figure 3.5 present the results of Traceback Accuracy as a function of the number of reported illnesses, K_ω. As expected, the Network Baseline performs better than the FDA Heuristic, since it makes use of a greater evidence base. However due to the high connective density characterizing each network, even after 150 have reported, the network baseline cannot significantly reduce the feasible source set. In other words, there are many stage $n = 1$ nodes that satisfy a connection to each node in the reported set $K_\omega = 150$, which we saw in Section 2.3.2 will be distributed across ~20 nodes (out of 25) on average. Neither method distinguishes between nodes in the feasible source set. When an analytical mechanism making use of network structure to distinguish between nodes within the feasible set is applied, the results are striking. Traceback Accuracy improves by almost 70% between the Network baseline and the method presented by B&H. Furthermore, the STT method performs significantly better than all existing approaches, demonstrating an improvement of >10% above B&H. These results suggest that the methods introduced in this Chapter might contribute substantially to improving outbreak investigation procedures.

We now apply the same comparison to the *Deterministic-4* network introduced Section 3.2.1. While the same comparative relationships hold between the FDA Heuristic, the Network Baseline, and STT, there is a patent difference between the detection performance of STT and B&H. While STT reaches 96% accuracy by $K_\omega = 150$, the existing method achieves 10% correct detections, performing slightly worse than even the Network Baseline. This means it is not able to accurately distinguish between nodes within the feasible set.

The lower performance exhibited by the state-of-the-art method in comparison with STT may be explained by the fact that the methodological approach of STT is tailored to the specific network problem investigated here. The Effective Distance method (B&H) is designed for general complex networks, directed or undirected, where possible paths travelled can differ markedly in length, by multiple orders of magnitude. Their approach leverages the observation that while a contaminant can travel a multitude of paths to any other node, the dynamics are dominated by the shortest paths; correspondingly, longer paths are penalized (Brockmann and Helbing 2013). In comparison, the STT algorithm is designed for directed, multi-partite networks where the length of all network paths from source node to contamination point are equal or close to the same number of steps, differing at most by the number of layers in the network. It leverages the observation that in multi-layer networks, the contamination will in fact travel across a multitude of paths; correspondingly, it considers the aggregate probability of all paths traveled from possible source to observed contamination point. This difference is exemplified by the *Deterministic* networks in which all paths statistically identical and no single path will dominate the contamination dynamics. The superior results demonstrated for STT in these network cases exemplify that it is the more appropriate method for the problem of source traceback on food distribution networks. It is also important to note that the *Deterministic-4* network is a stylized structure *designed* to exemplify the differences in the predictive approach taken by the two network-theoretical methods. As discussed in Section 2.2.4, real food distribution network structures demonstrate great degrees of heterogeneity in link distribution among other parameters, and the statistically identical structure typified by the *Deterministic* network is very unlikely to be observed in reality.


Traceback Accuracy Comparisons: Geometric Network

Traceback Accuracy Comparisons: Deterministic Network

Figure 3.5: Traceback Accuracy (TA) as a function of the number of illnesses K_ω for the **(a)** *Geometric-4* and **(b)** *Deterministic-4* networks. Results are based on 100 simulations in the baseline scenario. From bottom to top, TA with the FDA Heuristic is plotted in light green, with the Network Baseline in dark green, with B&H in red, and with the STT in blue.

Robustness of comparative behavior

We have performed extensive robustness analyses to verify whether that the comparative relationships exhibited in the two illustrative examples hold broadly. We have varied multiple structural parameters, including the distribution of flow volume across outgoing links from each node, the number of stages n, the number of nodes per stage v_n, and average degree μ. We have observed that the series of relationships FDA Heuristic < Network Baseline < B&H < STT holds almost universally, with a few fluctuations. Here we present two sets of analyses to demonstrate the consistency of results, applying the comparative framework to (i) a set of stylized network structures that vary significantly in their connective density and (ii) to the realistic network structures developed in Section 2.1.

We compare the two heuristics and two traceback methods across a set of 10 network structures defined by *Deterministic* out-degree distributions and 10 *Geometric* out-degree distributions, respectively. The networks considered in each figure vary in their average degree μ but are otherwise constructed according to identical variable specifications: Each network is a *square layered* network consisting of $n = 4$ stages of $v_n = 100$ nodes. The initial volume distribution is equal over the node set $V(1)$. Again, detection performance is based on 100 outbreak simulations generated according to the baseline scenario specifications. Figure 3.6 a and b presents the results obtained with the four methods applied to each series of 10 networks. Each figure plots Traceback Accuracy at a specific slice in time, $K_\omega = 150$ illnesses, and a function of average degree μ (thus, each integer value μ represents a network). As with the *Deterministic-4* network, STT significantly outperforms all existing methods for the set of *Deterministic* networks considered. For Geometric networks in which the majority of the flow from source to sink will be concentrated in a small number of high probability paths, the difference is much less pronounced while still significant. On average, STT performs 8% points better, ranging from 19% to 5% in all network cases except for $\mu = 8$, when STT dips 3% points below B&H. Fluctuations from the general trend are always possible for geometrically defined networks, which can exhibit extreme variability.

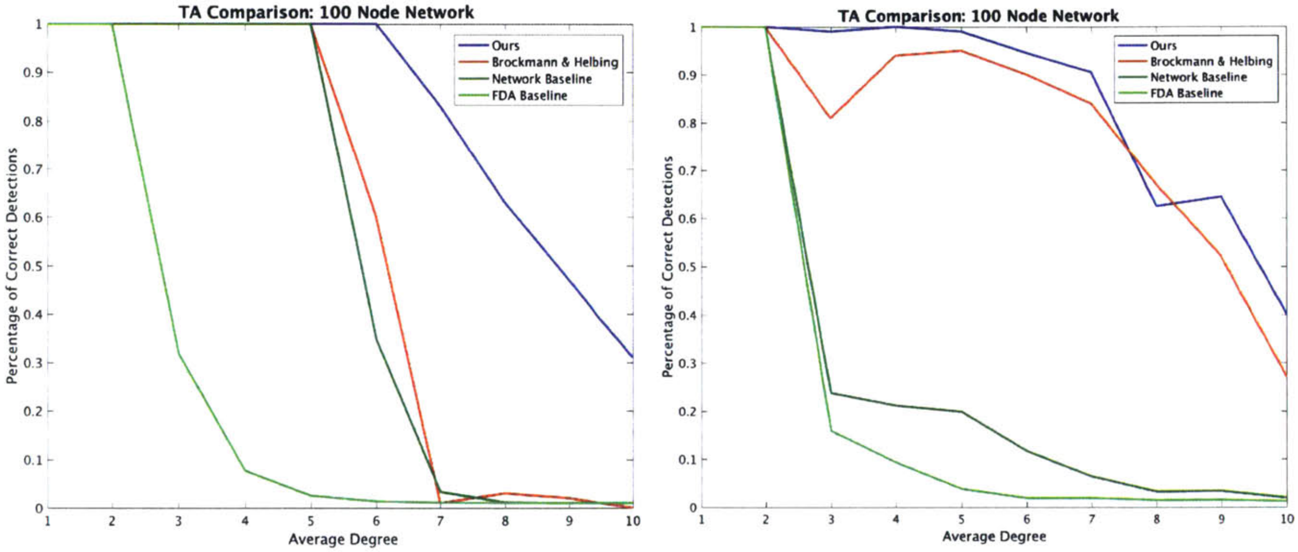


Figure 3.6: Traceback Accuracy (TA) at a specific slice in time, $K_\omega = 150$ illnesses, as a function of average degree μ for the **(a)** Tomato and **(b)** Lettuce networks. Results are based on 100 simulations in the baseline scenario. From bottom to top, TA with the FDA Heuristic is plotted in light green, with the Network Baseline in dark green, with B&H in red, and with the STT in blue.

We now demonstrate robustness of the comparative behavior in application to the realistic network structures developed in Section 2.1, for the performance evaluation framework described in Section 3.2. Figure 3.7 a and b presents the percentage of correct detection obtained with the four methods as a function of the number of illnesses K_ω. For both network cases, STT's detection performance is consistently best through the entire time course of the outbreak, closing out a 15% points better than B&H by K_ω=150 illnesses and outperforming the FDA Heuristic by an 80 – 85% margin.

3.3.3. Practical Implications

The comparison of the STT traceback method to heuristics and existing approaches suggests that the network-theoretical approach to the source localization problem can contribute substantially to the traceback investigation process. Current methods are able to identify the source in less than 35% of all identified outbreaks (Painter et al. 2013). If the source is unknown it is not possible for investigators to implement interventions to limit its spread, meaning that the majority of contamination events progress freely. Implementation of the methods introduced here may provide substantial benefits to emergency responders, helping them to identify the source successfully and efficiently, and importantly, enabling the implementation of interventions to avert illnesses before they occur. We will dig further into this difference and discuss the practical implications in Chapter 4, when we quantify the benefits resulting from using the traceback methodology to develop specific investigation interventions to identify the source and limit its spread.

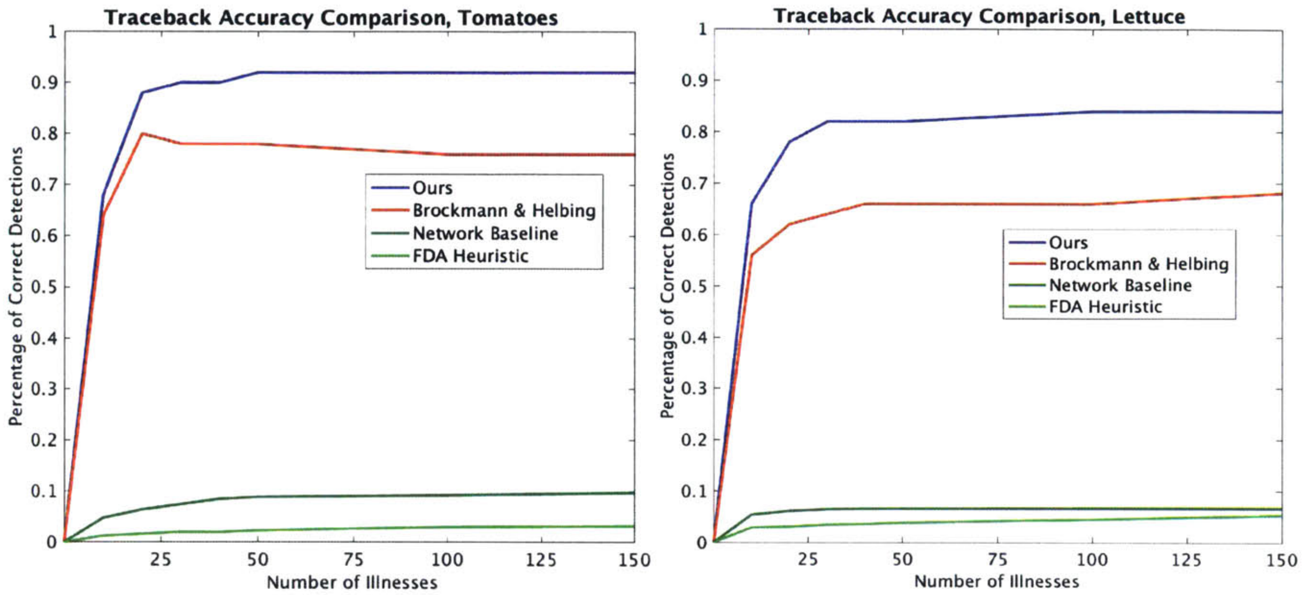


Figure 3.7: Traceback Accuracy (TA) as a function of the number of illnesses K_ω for the **(a)** Tomato and **(b)** Lettuce networks. Results are based on 100 simulations in the baseline scenario. From bottom to top, TA with the FDA Heuristic is plotted in light green, with the Network Baseline in dark green, with B&H in red, and with the STT in blue.

3.4. Conclusions

This chapter provides a first step towards validating the practical utility of the traceback methodology presented in this thesis in real-world scenarios. A generalizable modeling framework representing key structural and spatial features of real food supply networks has been developed and applied to model the distribution of two specific commodities in the US: tomatoes and lettuce. We have evaluated the performance of the traceback methodology applied to these two realistic structures, demonstrating high accuracy in both identifying and localizing the outbreak source. We have analyzed the results to provide an understanding of how the successful resolution of an outbreak depends on the structure of a network.

This chapter also develops a framework for quantifying benefits in comparison to existing approaches to foodborne disease outbreak source detection: to a practical baseline meant to demonstrate current methods applied in practice during outbreak investigations and to a best-in-class method presented in the literature. In results across a wide range of network structures and outbreak scenarios, our approach to traceback demonstrates significant improvements in accuracy. Theoretically, these results demonstrate the suitability of our specific methodological approach to the problem of localizing the source of foodborne disease outbreaks. Practically they suggest that our traceback methodology provides an effective framework for identifying the source of large-scale outbreaks of foodborne disease, and that it may contribute substantially to the traceback investigation process. In the following chapter, we explore the applicability of applying the method in real time, developing and implementing a decision-making tool for guiding investigators at the tactical level in making the most effective interventions to solve an investigation and stem impact on the public.

It is important to stress that these findings are derived from simulation results and as such are only illustrative. Future analyses will be necessary to determine the method's accuracy when applied to real food supply network data. This work is described in Chapter 5.

Chapter 4: Interventions

The performance evaluation presented in Chapter 3 demonstrates that the traceback methodology provides an effective framework for identifying the source of large-scale outbreaks of foodborne disease. The results suggest its implementation during an outbreak may provide substantial benefits. In this Chapter, we use the traceback methodology to develop a decision-making framework to guide investigators at the tactical level to make the most effective interventions to solve an investigation and stem impact on the public.

In the event of an outbreak, investigators' primary objective is to limit the number of illnesses. Illnesses are averted when the source is identified and the public is notified through a public service message, which can be combined with a recall and removal of the offending product from the supply chain. In the best cases, an initial investigation is successful in narrowing down the number of feasible sources to a few possibilities. Investigators are then deployed to these sites in order to gather records, observations, and perform sampling experiments in order to definitively identify the source. Often in practice, however, it may be possible only to narrow the problem down to a subset of facilities or a specific region. In some cases, such as during the spread of a particularly virulent or deadly outbreak strain, investigators may decide that it is more important to take measures to limit the spread before the source is singly identified. In these situations, a public service message regarding the status of the investigation may be issued, warning the public to avoid consuming products from specific brands or regional origin. In the worst cases, a category currently including over 65% of multi-state outbreaks occurring in the US, no leads on the source emerge in the investigation. In these cases, either no action is taken or an extremely broad message is issued that implicates an entire category of foods.

In deciding to deploy investigators or dispatch a public service message, investigators face difficult decisions. They must decide *whether* the accuracy of their assessments merits an action being taken, if so, then *when* it should be taken, and *what* it should include. The greatest challenge lies with determining *when*. Clearly, taking action earlier in the outbreak while the case development rate is highest will create the greatest benefit. At the same time, assessments regarding the source location made at an early stage in an outbreak's progression will, on average, be marked by greater uncertainty and a broader set of possibilities. If an untargeted message is issued implicating a large region or set of food products or categories, it will have major repercussions for all firms in that industry. For example, in the 2006 E.Coli spinach outbreak, all spinach in North America was pulled off the shelves while it took the authorities over a month to identify the origin of contamination (Seltzer et al. 2009). Tracing the spinach back to the county or even district of origin would have provided a huge impact both to consumers and the spinach industry. Furthermore, when a statement issued prematurely turns out to be incorrect, the indicted brand, product, or industry will suffer unmerited damages. This occurred in the 2011 German outbreak of E.coli in sprouts, when investigators wrongly implicated cucumbers produced by a Spanish produce cooperative, wiping out over a month's worth of production of that commodity and doing lasting damage to the reputation of the Spanish cucumber industry as a whole (a $2.54 million settlement was reached between the City of Hamburg, whose health officials made the mistaken implication, and the Spanish cooperative was reached in 2015). On the other hand, taking a conservative approach and waiting for more evidence to become available to pinpoint the source with greater certainty will inadvertently allow more cases of illness to proliferate. At the extreme, the epidemiological curve may well have died out by the time a confirmation is established and action taken, meaning that no illnesses will be averted.

Further complicating the difficulty of timing a decision is the criticality of each day in the outbreak's progression. This is exemplified by the example provided in Chapter 1 of various times at which action could have been taken to avert the spread of the 2006 spinach outbreak. Given the rate of case progression in that outbreak, five days of separation in the timing of an intervention would have resulted in a difference of 19 vs. 110 cases averted – around 8% vs. 70% of the total number of reported cases.

Standard protocol to guide the deployment of mitigation measures during an outbreak currently involves investigators consulting with colleagues at state public health agencies, university agricultural research institutions, and in industry to prioritize between sources. An importation limitation to these practices is that a systematized approach to identifying, evaluating, and deciding mitigation measures is not applied process. Sherri McGarry, the Foodborne Outbreak Coordinator at the FDA Headquarters has emphasized the need for scientifically sound approaches to guide investigation and control measures. Ms. McGarry has asserted that, "Any measure that will help to determine where we should focus our attention and give leads on the investigation would have a lot of application and utility for public health, and for business as the longer the outbreak the greater the impact on industry" (S. McGarry, personal communication, December 20, 2012).

In this Chapter, we propose, implement, and evaluate a systematic approach to establishing interventions that address the types of situations described. Mechanisms are developed to answer the questions *whether*, *when*, and *what*, deciding (i) when and where to **deploy investigators** and (ii) when to **message the public** implicating the likely outbreak source(s) and what locations it should include. The procedures involve investigators predetermining a desired accuracy level and allocating a non-monetary "budget" of resources to the investigation. In defining these strategies, we do not prescribe *how* a decision should be made; our purpose is to provide an objective framework that investigators can use to quantify and tradeoff their alternatives, making clear the benefits of taking (or not) certain actions.

The contributions of this chapter fall under four categories:

- We define the attributes accuracy, benefit to public health, and cost to regulators or industry as characterizing the performance of investigation interventions; these performance measures allow us to define an framework for enumerating, quantifying, and comparing intervention options.
- We propose mechanisms based on the traceback methodology of this thesis for deciding when and where to deploy investigators and when and what to message to the public given an allowable level of risk and the resources available.
- We quantify the potential benefits to public health possible if interventions based on these mechanisms are implemented, measured in terms of illnesses averted.
- We show from computational results that these methods demonstrate great potential to improve upon current methods in outbreak response, recommending whether, when, and with what to respond during an outbreak.

4.1. Intervention Performance Attributes

4.1.1. Overview

In the first section of this Chapter, we develop a decision-making framework and set of mechanisms for deploying investigators and dispatching messaging interventions to the public during an ongoing outbreak. The mechanisms require investigators to specify a desired accuracy target and an available resource budget. The procedures rely on the simulation-based traceback performance evaluation framework (Section 2.2), using the results of multiple simulation runs to quantify the expected improvement in traceback accuracy with each day's new information on illness reports. Based on the expected improvement in accuracy, a set of options is derived to indicate when and where to deploy investigators or messages for specific combinations of accuracy and resources. The decision maker will review the options and will make a decision by comparing the tradeoff between the expected accuracy, the cost to public health incurred by the number of new cases generated each day the outbreak progresses, and the cost to industry by implicating multiple firms.

In the following, we define intervention performance attributes and present the mechanisms. We apply the mechanisms to outbreak scenarios involving the Lettuce and Tomato networks and evaluate their potential to effectively recommend whether, when, and where to focus when responding to an outbreak. Implementing these methods also allows us to directly quantify the benefits of our methodology in comparison to existing methods used in outbreak investigations. Specifically, we quantify how many illnesses could have been avoided had a recall or public service announcement been made at the time of detection, given assumptions regarding the response time following the implication of the source(s).

4.1.2. Performance Attributes

We consider the following three performance attributes of interventions: accuracy, benefit to public health and cost to industry (i.e. what facilities or regions) to implicate. The specificity quantifies *what* action to take, accuracy pertains to the notion of *whether* it should be taken, and benefit to public health relates to *when* it should be carried out.

Specificity

This refers to an intervention's ability to focus on a small, bounded subset of top-ranked predictions. It is quantified by the number of facilities included in an intervention, S. In deploying investigators, it quantifies the number of facilities to sample, S_I. In dispatching a message to the public, it quantifies the number of facilities to implicate, S_M.

An intervention deploying investigators with a higher value S_I will have a larger cost to the central authority performing the investigation, as it will require a greater number of facilities to be inspected and sampled. An intervention to dispatch a message with a higher value S_M will have a larger cost to industry, as it will implicate a greater number of facilities when at most one will be culpable. Therefore, although the notion of "high specificity" may have a positive connotation in colloquial use, by this definition the connotation is negative.

Accuracy

This refers to an intervention's precision in correctly identifying the source within a specific number of top-ranked predictions, that is, of achieving a certain specificity. The Accuracy A is quantified by the probability that the true source s^* is ranked within the top S predictions, i.e. $A = P(R \leq S)$, where R is the random variable representing the Rank of the true source as defined in Section 2.2. The higher the accuracy, the greater the probability that the true source is identified.

Benefit to public health

This refers to an intervention's ability to mitigate the outbreak's impact on public health. It is measured by the number of illnesses averted by taking action to message the public not to consume the contaminated product (or set of products including the contaminated product). In interventions to deploy investigators, this action is taken after the S_I facilities have been inspected and sampled.

Illnesses Potentially Averted

To quantify the number of illnesses averted, we introduce a new metric, *Illnesses Potentially Averted* (*IPA*). *IPA* is defined for an outbreak O characterized by observed cases of illness $i \in O$. We define *IPA* as the number of cases *i* that had not consumed the contaminated product by the time a message to avoid that product would have reached them, for a messaging intervention strategy informed by the evidence provided by the first K_ω reported illnesses. To calculate the metric, it is necessary to describe a few important features of the process of implementing an intervention. Recall that cases *i* are recorded in the outbreak curve according to their time of illness onset t_ω. The eventual date that an intervention informed by the evidence provided by the set K_ω can realistically be implemented must account for two process delays: case reporting delay and message transmission delay. First, the date that each illness would become known to investigators and thus usable in traceback calculations is subject to a delay in case reporting which can range from a few days to 3 weeks, as described in Figure 1.1 and Appendix 1.1, the *Timeline for Case Reporting*. Second, after investigators use the available evidence to perform the traceback assessment, make predictions, and decide on an intervention, there is a delay between when the message is announced and when it reaches members of the consuming public. This may include the time it takes for the message to spread through various media outlets, or for food providers to remove the item(s) from circulation. In determining *IPA*, we account for the case reporting delay and the messaging delay. We assume that the time to perform the traceback assessment, make predictions, and decide on an intervention is instantaneous.

IPA is calculated for O using the outbreak simulation model (Section 2.2), with a few additions. First, we introduce the following notation:

- $t_{c,i}$, the date that *i* consumes the contaminated product
- t_i, the date that symptoms begin for *i*, which is the date recorded in $E(t)$
- Θ_R, a random variable representing the case reporting delay
- $t_{r,i}$, the date that *i* is ultimately reported to investigators
- τ_{K_ω}, the time that the K_ω th illnesses would be reported
- Θ_M, a random variable representing the message propagation delay
- $t_{m,i}$, the date the message would reach *i*

Dates of consumption and illness onset $t_{c,i}$ and t_i are determined according to the process represented in the original simulation model, described in Section 2.2. To determine the reporting times $t_{r,i}$, a reporting delay θ_R is sampled from Θ_R and added to the date of illness onset t_i. The time τ_{K_ω} can then be found; set equal the time $t_{r,i}$ associated with the K_ω th reporting illnesses.

Informed by the evidence provided by K_ω, investigators will perform the traceback assessment, make predictions, and decide on a messaging intervention. Since we assume this decision is made instantaneously, τ_{K_ω} is stored as the *Date of Investigation Response*. The message is propagated, reaching case *i* at $t_{m,i}$ after a random propagation delay θ_M sampled from Θ_M. Finally, we can find *IPA* for outbreak O as the number of cases *i* reached by the message before consuming the contaminated product, i.e.

$$IPA = \sum_{\{i \in O | t_{m,i} \le t_{c,i}\}} i \qquad (4.1)$$

The critical assumptions of this approach are that the message will eventually reach all retailers and consumers, who once reached, will react immediately by removing or discarding the contaminated product. These assumptions are not entirely realistic, since retailers and consumers in more isolated parts of the country may not be reached by the message. However since the industry-led recall of the contaminated product following its message will contribute to the reduction of that product in the supply chain, it will become increasingly less possible that the contaminated commodity remains in circulation, meaning that each consumer or food seller will ultimately be *affected* by the intervention itself. The uncertainty in the message propagation delay variable Θ_M will account, in part, for the variability in the time the intervention takes to achieve this *affect*.

We also note that the quantification of *IPA* is limited to reported illnesses; illnesses that go unreported are not considered. Due to underreporting or unknown agents, these numbers are potentially much higher, with estimates that for every reported illness, between 25 – 100 go unreported (Batz 2005, Batz and Morris 2012, Scallan et al. 2011). The true public health benefit is thus larger than indicated by *IPA*.

4.2. Mechanisms for Interventions

In this section, we develop several mechanisms for outbreak interventions that define inputs provided by investigators and determine set of feasible intervention strategies satisfying the input conditions. In defining these mechanisms, we do not prescribe *how* investigators should compare between feasible strategies to make a decision, such as by assigning values to the "importance" of each performance attribute in order derive a solution that "optimally" trades off between them. Rather, our intention is to provide a systematic framework that investigators can use to (i) identify a set of possible intervention options, (ii) quantify the costs and benefits of each according to the performance measures, and (iii) compare the options on an objective basis.

4.2.1. Investigator Deployment Mechanism

We first define a mechanism for deciding where and when to deploy investigators to sample suspect source facilities. Under this mechanism, the investigation decision-makers, which we refer to as the "investigation task force," submit their preferred accuracy target $A^{\min}$ and resource budget $S_I^{\max}$, determining the maximum number of facilities that they can afford to sample. With regard to specifying input values, we note that the accuracy target is flexible and defines an "acceptable" level of risk investigators are comfortable taking on in implementing an intervention. While the resource budget may be inflexible, set by a specific finite budget, it may also be negotiable. For example, emergency funds might be allocated to respond to an outbreak strain causing greater "impact" measured by the severity of the presenting cases.

The accuracy target and resource budget imply the constraint $P(R \le S_I) \ge A^{\min}$, for $S_I \le S_I^{\max}$. Given the inputs, the decision modeler, who we refer to as the "analyst," forms a set of feasible intervention options or strategies $\phi \in \Phi_I$, each describing a unique combination of accuracy and resource budget parameters satisfying this constraint. The analyst then uses the simulation-based traceback performance evaluation framework (Section 2.2) to quantify the expected public health benefit of each strategy, defined as the mean number of illnesses potentially averted $\overline{IPA}$. First, the PMF for Rank R, $P_R(r)$, is formed from the results of the traceback algorithm applied to a set of simulated outbreak events $O \in \mathrm{O}$. $P_R(r)$ is used to determine $K_\omega^{\min}$, the minimum amount of evidence, i.e. the smallest number of reported illnesses K_ω, necessary to achieve each strategy $\phi \in \Phi_I$. Finally, the expected public health benefit $\overline{IPA}$ is calculated by averaging IPA over all outbreak events. Recall that IPA is a metric defined for an individual outbreak event O and evidence level K_ω. $\overline{IPA}$ is therefore also a function of the evidence level, and will be identical for each strategy with the same revealed value $K_\omega^{\min}$.

The investigation task force reviews the intervention strategies, comparing the alternatives on the basis of their accuracy (quantified by A), cost to the central authority performing the investigation (quantified by resources required S_l), and benefit to public health (quantified by $\overline{IPA}$). The task force chooses a strategy; establishing *what* action will be taken, i.e. deploy investigators to S_l sources, and *when* the action will be taken, i.e. after waiting for K_ω illnesses to report. This mechanism does not prescribe how the task force should arrive at a decision, though the first step should be to quantify and compare the tradeoffs between the intervention strategies. For example, it would be useful to determine the increase in the cost to public health incurred by waiting for a more stringent accuracy target to be achievable at a given resource budget, or alternatively, waiting until a desired level of risk is achievable at a lower resource budget.

Finally, the analyst shifts focus from the results of the set of O simulated/hypothetical outbreak events and back to the real data from the ongoing outbreak event. The final step is carried out once the K_ω th illness has been reported and the analyst can determine the realization of the strategy, identifying *where*/which S_l facilities investigators are deployed to. The traceback algorithm is applied to determine which facilities occupy the first S_l positions in the resulting ordered ranking. The intervention is successful if the true source s^* is identified within this set. Follow-on action should be taken to warn the public to avoid the contaminated product.

Mechanism 1: Investigator Deployment Mechanism

- The investigation task force submits their preferred accuracy target $A^{\min}$ and maximum number of facilities to deploy investigators to $S_l^{\max}$, determined by the available resource budget
- The analyst forms a set of feasible intervention strategies $\phi \in \Phi_l$ for combinations of accuracy and resource budget parameters, satisfying the constraint $P(R \le S_l) \ge A^{\min}$ for $S_l \le S_l^{\max}$.
- The analyst quantifies the expected benefit to public health for each intervention using the simulation-based traceback performance evaluation framework (Section 3.2)
    - The PMF for Rank R, $P_R(r)$, is formed from the results of the traceback algorithm applied to a set of simulated outbreak events $O \in \mathrm{O}$.
    - $P_R(r)$ is used to determine $K_\omega^{\min}$, the minimum amount of evidence necessary to achieve each strategy $\phi \in \Phi_l$
    - The mean number of illnesses potentially averted $\overline{IPA}$ is defined for each strategy according to $K_\omega^{\min}$
- The investigation task force reviews the strategies, comparing the alternatives on the basis of their accuracy (quantified by A), cost to the central authority performing the investigation (quantified by resources required S_l), and benefit to public health (quantified by $\overline{IPA}$), and decides on an intervention.
- Once the once the K_ω th illness has been reported, the analyst determines the realization of the strategy, identifying *where*/which S_l facilities investigators are deployed to.
    - The analyst shifts focus from the results of the set of O simulated/hypothetical outbreak events and back to the real data from the ongoing outbreak event.
    - The traceback algorithm is applied to determine which facilities occupy the first S_l positions in the resulting ordered ranking.

4.2.2. Public Service Message Dispatching Mechanism

We now propose a mechanism for deciding when to issue a public service message implicating the likely outbreak source(s) and what facilities it should include. The mechanism applies the same approach as Mechanism 1, the difference being the meaning of input and output terms. As in Mechanism 1, investigators provide their preferred accuracy target $A^{\min}$, but instead of an internal resource budget, they submit a maximum allowable number of facilities to implicate in the message $S_M^{\max}$. If $S_M^{\max} > 1$, this term defines an "acceptable" limit on the cost to industry incurred by the damage caused by an unmerited attribution since most one facility will truly be the source.

As for mechanism 1, strategies satisfying the input constraints are formed and quantified according to their expected benefit to public health and the investigation task force reviews the strategies in order to decide on an intervention strategy $\phi \in \Phi_I$ carried out at interval K_ω. In this case, however, alternatives are compared on the basis of their damage to industry (quantified by number of facilities implicated S_M) in addition to their accuracy (quantified by A) and benefit to public health (quantified by $\overline{IPA}$). The task force chooses a strategy; establishing *what* action will be taken, i.e. implicate S_M facilities in the message, and *when* the action will be taken, i.e. after waiting for K_ω illnesses to report.

Finally, once the K_ω th illness has been reported, the analyst determines the realization of the strategy, identifying *which* S_M facilities to implicate. The traceback algorithm is applied to determine which facilities occupy the first S_M positions in the resulting ordered ranking. A messaging intervention is successful if the true source s^* is within this set, meaning that the public has been warned to avoid the contaminated product.

Mechanism 2: Public Service Message Dispatching Mechanism

- The investigation task force submits their preferred accuracy target $A^{\min}$ and maximum number of facilities to implicate in message $S_M^{\max}$, quantifying an "acceptable" limit on the damage/cost to industry
- The analyst forms a set of feasible intervention strategies $\phi \in \Phi_I$ for combinations of accuracy and number of facilities implicated, satisfying the constraint $P\left(R \le S_M\right) \ge A^{\min}$ for $S_M \le S_M^{\max}$.
- The analyst quantifies the expected benefit to public health for each intervention using the simulation-based traceback performance evaluation framework (Section 3.2)
    - The PMF for Rank R, $P_R\left(r\right)$, is formed from the results of the traceback algorithm applied to a set of simulated outbreak events $O \in \mathrm{O}$.
    - $P_R\left(r\right)$ is used to determine $K_\omega^{\min}$, the minimum amount of evidence necessary to achieve each strategy $\phi \in \Phi_I$
    - The mean number of illnesses potentially averted $\overline{IPA}$ is defined for each strategy according to $K_\omega^{\min}$
- The investigation task force reviews the strategies, comparing the alternatives on the basis of their accuracy (quantified by A), specificity (quantified by facilities implicated S_I), and benefit to public health (quantified by $\overline{IPA}$), and decides on an intervention.
- Once the once the K_ω th illness has been reported, the analyst determines the realization of the strategy, identifying *which* S_M facilities to implicate in message.
    - The analyst shifts focus from the results of the set of O simulated/hypothetical outbreak events and

back to the real data from the ongoing outbreak event.

- The traceback algorithm is applied to determine which facilities occupy the first S_M positions in the resulting ordered ranking.

4.3. Implementation

In the following we implement the mechanisms to the Tomato and Lettuce networks introduced in Chapter 3. Assuming reasonable values for input constraints, we present a set of strategies prescribed by the mechanisms. The impact of the interventions is quantified according to the performance measures. Finally, we compare this approach to current methods in outbreak investigation, quantifying the benefits of a systematic approach to developing interventions, and moreover, one based on the novel approach to traceback introduced in this thesis.

4.3.1. Modeling Assumptions

First, we report on our input parameter choices. We consider an accuracy target of $A^{\min} = 90\%$. This can be interpreted as accepting the risk that 1 out of every 10 interventions implemented is unsuccessful. Investigation resources were limited to $S_I^{\max} = 5$ independent investigations to reflect the very limited resources available in the federal outbreak response budget. The maximum number of facilities to implicate in a message was also limited to $S_M^{\max} = 5$, implying it would be too damaging to industry to wrongfully implicate more than 5 brands. To satisfy the bounds set by these thresholds, we consider strategies at combinations of $A = 90\%$ and 95% , and S_I , $S_M \le 5$.

To calculate $\overline{IPA}$, we apply the traceback performance evaluation framework to the *E.coli* outbreak scenario and simulation set considered in Section 3.2, leading to 374 and 462 reports of illness on average for the Tomato and Lettuce cases. We model the case reporting delay using a Weibull random variable $R \sim \text{Weibull}(\lambda,\kappa)$ with scale parameter $\lambda = 7$ and shape parameter $\kappa = 1.5$, a right-skewed distribution peaked at 7 days with a long right tail. We model the message propagation delay M as a normal random variable with a mean of $u_M = 0.5$, or half of a week, and with variance $\sigma_M^2 = \frac{1}{36}$, so that most of its density occurs between 0 and 1 week.

4.3.2. Results

We first discuss what type of strategy should be adopted for each network case by considering qualitative results. Figure 4.1 a and b plot traceback accuracy against $\overline{IPA}$, the mean number of cases potentially averted , as a function of the available evidence K_ω . As we saw in Section 3.2, the traceback methodology converges quickly, almost reaching its peak accuracy with the evidence from between 20 – 30 cases of illness. At the same time, $\overline{IPA}$ falls quickly towards the beginning of the outbreak, reflecting the bell shaped curve of the average epidemic that increases rapidly and then decreases almost as rapidly. This comparison enables an important observation for both cases: compared with interventions implemented late in the outbreak's progression, interventions implemented early on will achieve a much greater benefit to public health without taking on greater risk. In the following, we therefore only present interventions achieved at available evidence levels $K_\omega \le 50$.

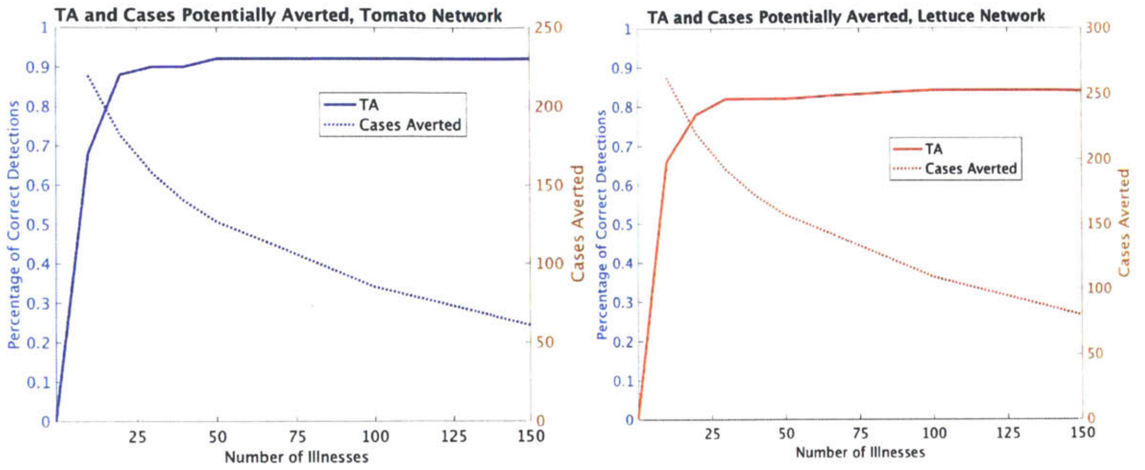
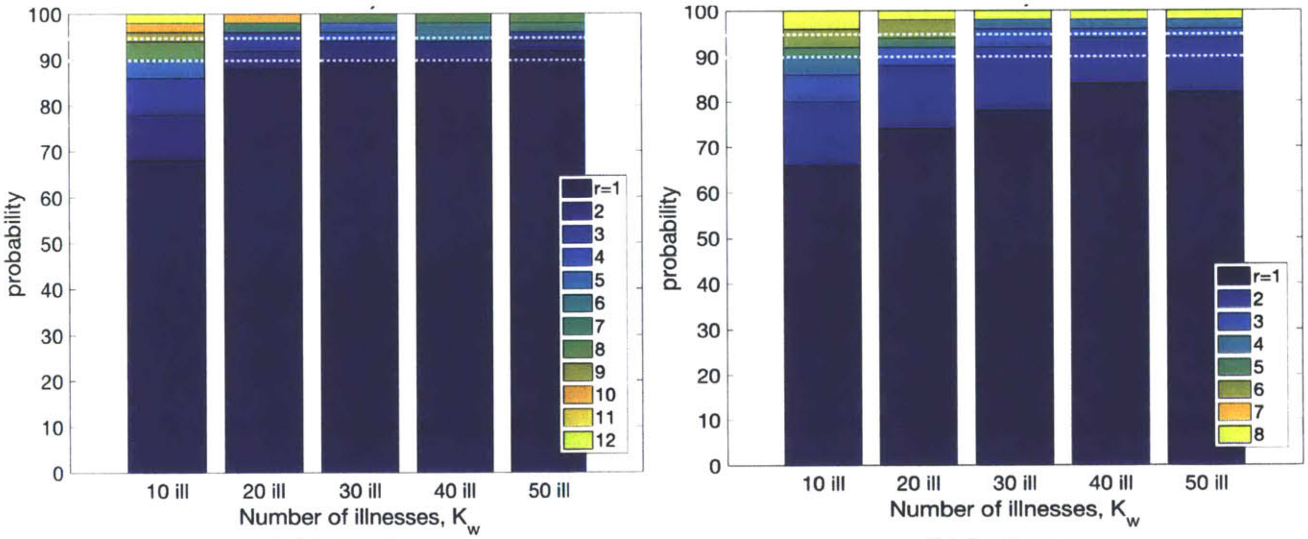


Figure 4.1. Traceback Accuracy and mean number of cases potentially averted $\overline{IPA}$ as a function of the number of illnesses K_ω for **(a)** the Tomato network and **(b)** the Lettuce network. The comparison qualitatively illustrates that interventions implemented early on can achieve a much greater benefit to public health without taking on greater risk.

To illustrate how the mechanisms are implemented, the Rank PMF $P_R(r)$ formed from the results of simulation and used to calculate $K_\omega^{\min}$ is demonstrated in Figure 4.2 a and b. The white dashed lines indicate accuracy levels of 90% and 95% being met $P_R(r)$ can also be used to quantify alternative strategies, such as determining how many facilities would need to be sampled / implicated to implement a strategy with 100% accuracy, or what strategies could be postulated after only 10 illnesses have reported.



(a) Tomatoes **(b) Lettuce**

Figure 4.2. Rank PMF $P_R(r)$ formed from the results of simulation and used to calculate $K_\omega^{\min}$ for the Investigator Deployment and Message Dispatching Mechanisms.

We now report the intervention options presented by the mechanisms. Table 4.1 shows the minimum amount of evidence necessary to achieve the constraints (i.e. the smallest number of reported illnesses, $K_\omega^{\min}$) and the expected benefit to public health (i.e. the mean number of illnesses averted, $\overline{IPA}$) for each strategy. To facilitate comparisons between the two network cases, the benefit to public health is also represented as a percentage of the total number of cases of illnesses. A dash denotes strategies that cannot be achieved at any level of evidence. Figures 4.3a and b visualize the expected benefit to public health. Each figure plots the average results of the total epidemic curve with no intervention, overlaid by the diminished epidemic curves that could result from successful interventions implemented at the evidence level $K_\omega^{\min}$ corresponding to each strategy.

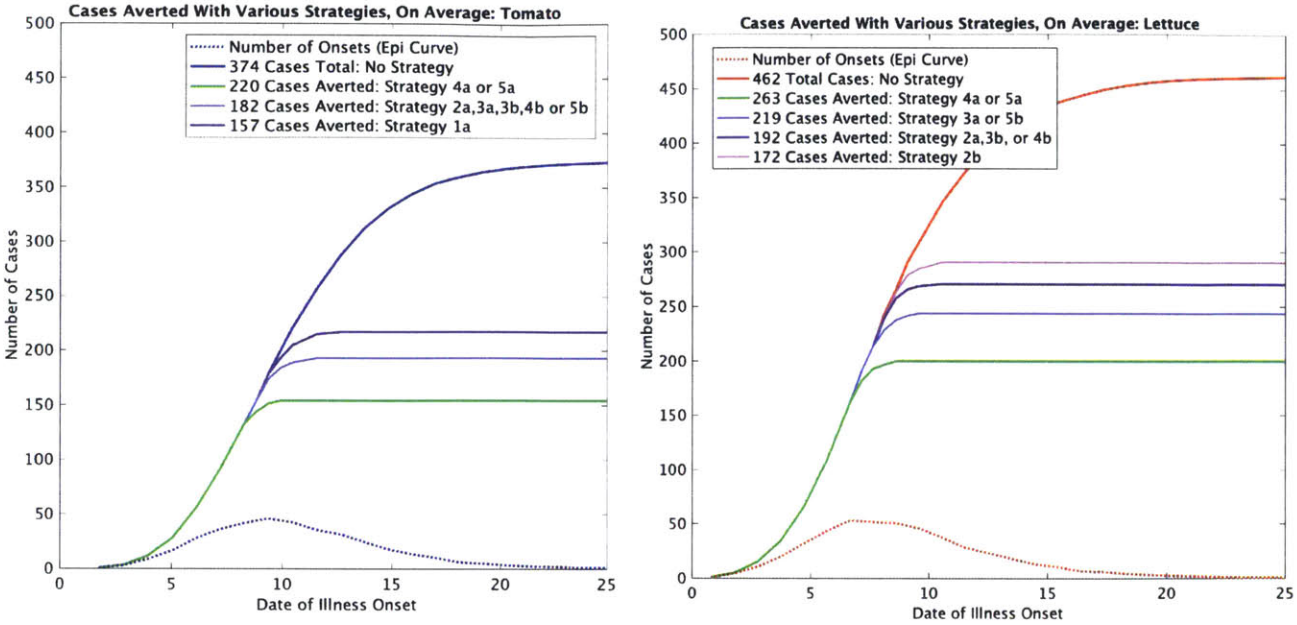
Intervention				Implementation: Tomatoes			Implementation: Lettuce		
Strategy	Facilities Sampled, S_I	Facilities Implicated, S_M	Accuracy Target, A	Evidence required, $K_\omega^{\min}$	Mean Cases Averted, $\overline{IPA}$	$\%\overline{IPA}$	Evidence required, $K_\omega^{\min}$	Mean Cases Averted, $\overline{IPA}$	$\%\overline{IPA}$
1a	Sample 1	Implicate 1	90%	30	157	42%	—	—	—
1b	Sample 1	Implicate 1	95%	—	—	—	—	—	—
2a	Sample 2	Implicate 2	90%	20	182	50%	30	192	42%
2b	Sample 2	Implicate 2	95%	30	157	42%	40	172	37%
3a	Sample 3	Implicate 3	90%	20	182	50%	20	219	48%
3b	Sample 3	Implicate 3	95%	20	182	50%	30	192	42%
4a	Sample 4	Implicate 4	90%	10	220	60%	10	263	57%
4b	Sample 4	Implicate 4	95%	20	182	50%	30	192	42%
5a	Sample 5	Implicate 5	90%	10	220	60%	10	263	57%
5b	Sample 5	Implicate 5	95%	20	182	50%	20	219	48%

Table 4.1. Results of Investigator Deployment and Message Dispatching Mechanisms for an outbreak of *E.coli* in Tomatoes and Lettuce.

We first note that almost all strategies are "achieved," meaning that there exists a value K_ω at which $P(R \le S) \ge A$. All but the most demanding constraint combination are achieved for the Tomato network and all but the two strategies requiring only 1 facility to be sampled/implicated are achieved for the Lettuce network. All other strategies suggest major benefits upon implementation, ranging from between at least 37 – 40% and at most 57 – 60% illnesses averted. The range of benefits is so similar for the two cases because of the parallel behavior of traceback accuracy, rising quickly and almost reaching its peak value at the same evidence level by around which rises quickly and almost reaches its peak value by $K_\omega =$ 30, as seen in Figure 4.1 above.

To choose a strategy, the investigation task force would compare the alternatives on the basis of their preferred risk criteria, resources available or cost to industry, and the expected benefit to public health. Figures 4.3 a and b help to visualize specific tradeoffs between intervention options. For example, 95 out of every 100 interventions implemented in the Lettuce case will be successful, averting on average 219 cases for a cost of 5 investigations or implications. To achieve a cost reduction by 2, 27 fewer cases would be averted, and to achieve a cost reduction by 3, 47 fewer cases would be averted. The meaning of these differences and the implications for the decision-making process will ultimately be determined by the severity of the disease. One can imagine that a difference of 27 illnesses will drive a more urgent response if a large fraction of these cases are expected to present with life-threatening complications or mortality.

We also note that the quantification of $\overline{IPA}$ is limited to reported illnesses; illnesses that go unreported are not considered. Therefore, the true public health benefit is larger than indicated by the figures in Table 4.1 Even without this multiplier, it is clear that great gains in public health can be achieved by improved traceback processes.



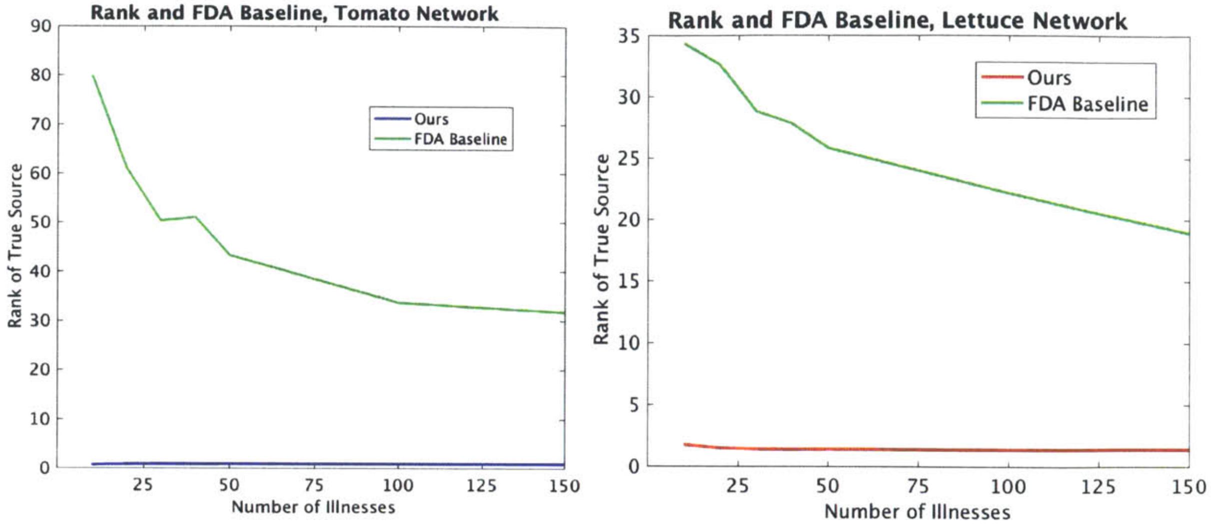
(a) Tomatoes **(b) Lettuce**

Figure 4.3. Visualizing tradeoffs and benefits of various investigation strategies implemented to mitigate the spread of an outbreak of *E. coli*.

4.4. Comparison to Current Investigation Methods

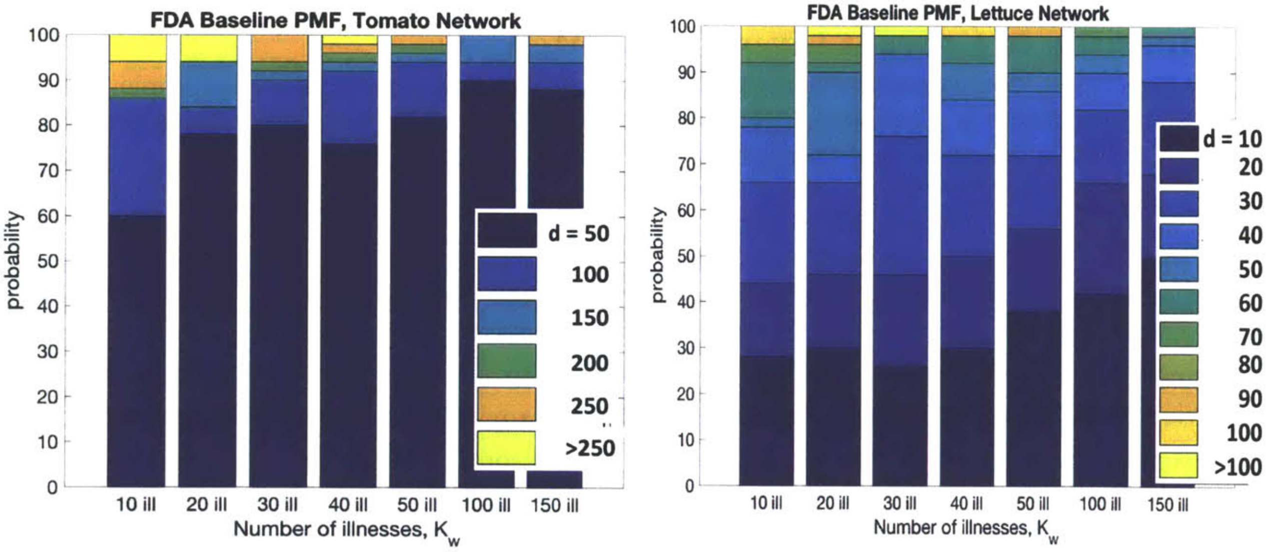
To quantify benefits in comparison with current methods in outbreak investigation, we implement the FDA Heuristic introduced in Section 3.3. The heuristic was developed from discussions with FDA investigators. It models the process of "triangulation," or tracing back the distribution paths of products from several locations to determine if there is a common point of convergence in the supply chain. Ideally, we would evaluate the heuristic in combination with the decision-making framework developed here, quantify the resulting performance attributes, and compare with the results achieved with the traceback methodology presented above. However, no strategies satisfying the pre-specified input constraints are achievable with the FDA Heuristic. To see this, in Figures 4.4a and 4.4b we plot the average Number of Feasible Sources $\overline{F}$ converged on by the FDA Heuristic against the Rank of the True Source metric $\overline{R}$ found with the traceback methodology, both as a function of number of illnesses K_ω. The FDA Heuristic is able to narrow down the number of possible sources to around 30 possibilities in the Tomato network case and 20 possibilities in the Lettuce case, and this is *on average*; the range of possible values is much greater. Clearly, strategies requiring a specificity of $S_I^{\max}, S_M^{\max} \leq 5$ are not feasible.

Therefore, to quantify benefits, we measure the equivalent specificity S_F (i.e. the investigative resources or cost to industry), required for the FDA Heuristic to achieve the same benefits to public health (i.e. an intervention implemented at the same evidence level $K_\omega^{\min}$) at the same minimum allowable risk (i.e. an accuracy target of $A^{\min} = 90\%$) achieved by the traceback methodology – mechanism combination. This is equivalent to finding the value of S_F satisfying $P(R \leq S_F) \geq A$ for the value $K_\omega^{\min}$ corresponding to each strategy. S_F can determined by inspecting the PMF for F, shown in Figure 4.5.



(a) Tomatoes (b) Lettuce

Figure 4.4. Plots of the average **(a)** Number of Feasible Sources $\overline{F}$ converged on by the FDA Heuristic and **(b)** Rank of the True Source metric $\overline{R}$ found with the traceback methodology, as a function of number of illnesses K_{ω} .



(a) Tomatoes (b) Lettuce

Figure 4.5. PMF for F , Number of Feasible Sources converged on by the FDA, formed from the results of simulation and used to quantify the specificity S_F .

Tables 4.2 and 4.3 report the results of this comparison, showing the specificity S_F required by the FDA Heuristic to achieve the same accuracy A, benefit to public health $\% \overline{IPA}$, and specificity S achieved by our methodology. Clearly, the traceback methodology developed in this thesis outperforms the FDA heuristic by a significant margin. To identify the source with 90% accuracy and avert the same number of illnesses, the minimum specificity S_F required by the FDA Heuristic ranges from 40 – 100. When interpreted according to the perspective of the mitigation measures, these values are impossibly high. In terms of investigator deployment interventions, it would not be feasible to deploy investigators to this many facilities within the timeframe required by emergency response measures. For messaging interventions, implicating this many facilities in a public service message would essentially represent condemning the entire tomato or lettuce industry.

Thus, for the outbreak scenario and distribution networks considered, it would not be possible to implement *successful* mitigation measures according to current methods in outbreak response. Furthermore, these results suggest that the large benefits to public health that come from *successful* interventions are possible only with the traceback methodology developed in this thesis combined with the decision-making tool of this Chapter.

Strategy	Accuracy Target, A	Public Health Benefit, $\% \overline{IPA}$	Specificity S (Ours)	Specificity S_F (FDA Heuristic)
1a	90%	42%	1	100
2a	90%	50%	2	150
3a	90%	50%	3	150
4a	90%	60%	4	200
5a	90%	60%	5	200

Table 4.2. Results of Investigator Deployment and Message Dispatching Mechanisms for an outbreak of *E.coli* in Tomatoes.

Strategy	Accuracy Target, A	Public Health Benefit, $\% \overline{IPA}$	Specificity S (Ours)	Specificity S_F (FDA Heuristic)
1a	90%	—	—	—
2a	90%	42%	2	40
3a	90%	48%	3	50
4a	90%	57%	4	60
5a	90%	57%	5	60

Table 4.3. Results of Investigator Deployment and Message Dispatching Mechanisms for an outbreak of *E.coli* in Lettuce.

4.5. Conclusions

In this Chapter, we developed, implemented, and evaluated mechanisms for investigation interventions that build upon the ranked output of the traceback methodology. These interventions can help investigators to determine *when* and to *what* facilities to deploy investigators to or to implicate in a message dispatched to the public. While specific benefits are determined by the particular outbreak scenario, general conclusions can be drawn regarding the potential impact of the mitigation-based approach to developing and implementing interventions. In particular, the computational results presented here suggest that benefits can be measured on four important dimensions:

1. If interventions based on these approaches are implemented, the true source can be identified within specific number of high probability candidates with specific accuracy.
2. This identification is possible early enough in an outbreak's progression that if quick action is taken, a substantial fraction of the illnesses that would ultimately result can be averted: ~40 – 60% in simulated outbreaks.
3. All of these results can be achieved with low demands on investigational resources (if investigators are deployed) and/or cost to industry (if a message is dispatched to the public).

4. The decision-making framework combined with the traceback methodology outperforms heuristics meant to model current methods in traceback investigation by a significant margin, requiring a fraction of resources necessary to achieve the same level of benefits.

It is important to stress that these findings are derived from simulation results and as such are only illustrative; live use of these techniques has yet to occur and may demonstrate features of the real problem inadvertently omitted from the modeling. The network models underlying the analysis are neither complete nor perfectly representative of real distribution structure or complete. More fundamentally, the accuracy of the system will ultimately depend on the fidelity of the underlying network model and the outbreak illness data available at the time of an event. This data is always better in simulation results than at the time of an event, even when realistic scenarios are recreated. Still, the results suggest that when applied to real outbreak investigations, the decision-making framework will present major improvements to the traceback process, resolving outbreaks that would not be solved by current methods and preventing many illnesses. At its best, it can identify the source definitively. At its worst, it can still narrow down the problem to a feasible set of sources, providing investigators with guidance at a tactical level.

Chapter 5: Implementation: Modeling the Food Supply Network

The traceback framework developed in this thesis presents a viable approach to improving the capacity to effectively and rapidly identify the source of outbreaks of foodborne disease. In Chapters 3 and 4, we quantified potential benefits of this methodology to public health, industry, and investigators by improving the speed, effectiveness, and accuracy of traceback investigations on a large scale. Here, we describe how our methodology might be implemented to form a *holistic system* for rapid traceback of outbreak events.

Vital to using the traceback framework during actual contamination events is a real-world supply chain network model for any food involved in a contamination event. To implement the traceback methodology in an emergency, a database of network models for various food types would need to be constructed in advance so that the methodology is ready to launch immediately. However acquiring and organizing the necessary food system data presents multiple practical challenges. Due to these challenges, it would be opportune to collect the minimal data necessary to achieve high traceback performance without oversupplying it. Clearly, the performance of the combined model-and-method approach will ultimately depend on the properties of the underlying network, including the granularity of the model and the representativeness of the data informing the model. In this chapter, we seek to understand the parameters of the "sufficient" network data. We propose multiple approaches to modeling the supply chain network, each at a different level of detail or granularity. For each approach, we suggest a means to collect the necessary data and then examine (i) the feasibility of collecting this data and (ii) the potential traceback accuracy achieved when implementing the model together with the traceback method. We start by revisiting the basic requirements of the supply chain network model and outlining the data needs of the "idealized" system and considering the feasibility of implementation. We examine the practical challengs associated with collecting this data; namely, the feasibility of capturing and storing it, and the high compliance burden doing so would pose to private enterprise. We then propose four approaches to forming the model at different levels of granularity, considering the accuracy, feasibility, and ready implementability of each. On the basis of the last approach presented, in third section we a recommend ready-to-implement system for real-time source detection.

The contributions of this chapter fall in four categories:

- We specify the requirements of the underlying system-wide supply chain network model and propose four approaches for modeling the structure meeting the necessary requirements
- We examine the potential accuracy of each alternative, considering in particular the level of detail necessary to achieve high traceback performance without oversupplying it
- We suggest a means to collect the necessary data and discuss the feasibility of its implementation
- On the basis of these analyses, we recommend a combined model-and-method approach that would form a ready-to-implement system for real-time source detection and suggest next steps in its evaluation and implementation.

Section 5.1 outlines the "idealized" network model and considers the challenges associated with its implementation. Section 5.2 proposes four approaches to forming the model at different levels of granularity, considering the accuracy, feasibility, and ready implementability of each. On the basis of the last approach presented, in Section 5.3 recommends a ready-to-implement system for real-time source detection. Section 5.4 concludes.

5.1. Network Models for Implementation

In this section we revisit the requirements of the supply chain network model and outline the idealized data needs of the system and considering the feasibility of implementation. As discussed in Section 2.1.1, the network model must take a systemic view for a given commodity, since the national food supply is a coherent system. This system consists of individual *supply chain actors* of various kinds, e.g. growers, processors, distributors, and retailers, sometimes organized under parent companies or *businesses*, and markets that organize the movement of goods between actors, e.g. retailing, wholesaling, and direct sales. The network model must capture the flows of goods between actors and between businesses, documenting the many possible paths from point of production through processing, production, transportation and distribution, to point of sale at retail.

5.1.1. Idealized solution

The network-specific inputs required by the traceback framework, summarized in Table 2.1, are the topology of the network, characterized by (i) the **nodes** and (ii) the **links and flows**; (iii) the geographical **location** of and **distance** between the nodes; and (iv) the **storage time delays** at each node. We now provide an overview of the ***idealized*** data for each of these input categories, which would form the most complete picture of the network, which is summarized in the third column in Table 5.1.

Nodes

In the idealized view, **nodes** would represent each individual supply chain actor engaged in preparing food for consumption, for each stage of the supply chain. Precise geographical coordinates would document the **location** of each node. The most accurate **distance** between any pair of nodes is the shortest path road network distance between them. With the geographical coordinates, this can be calculated trivially using GIS software.

Links and Flows

According to the network model framework in the traceback methodology, **links and flows** are equal to the total volume of commodity sent from supply chain actor u in one stage to each adjacent actor v in the subsequent stage, as a fraction of total volume handled by u per unit time. Here, we define the unit time to be the *temporal period of analysis* τ_p, which could be a week, growing season, or year. It is important to emphasize that because of this feature, the network represents a probabilistic picture of all connectivity over a fixed period of time rather than representing transactions as they happen. As long as all *possible* links occurring within τ_p are logged, including contracted relationships, spot-market relationships, and possible but unknown relationships, this probabilistic picture accounts for the inherent stochasticity in the market. Clearly, the smaller the period τ_p of connections the network structure represents, the more accurate the model will be.

Assembling link and flow information for the idealized view would require documenting the existence of every pair of trading supply chain actors and the total magnitude of trade over each link. This information could be gathered exhaustively by collecting and reviewing all transaction data over a fixed period of time. Importantly, it would need to be updated as frequently as τ_p.

Storage Time Delay

The **storage time delay** for a supply chain node is equal to the difference between receiving and dispatching times of commodity handled by that node. In the idealized case, this would be documented for each supply chain node u, rather than taken as an average across all nodes of stage n as implemented in the outbreak scenarios and traceback analyses conducted in Chapters 2 and 3. In the exhaustive case, this term could be determined by collecting a representative sample of transaction data and analyzing the facility logistics to determine a representative distribution of the storage times; the sample should be large enough to account for outliers.

Traceback Model Input	Definition	Description	Data To Collect
Nodes	Supply chain nodes are actors *u* in stage *n,* for all stages $n \in 1,...,N$	Nodes represent supply chain actors engaged in preparing food for consumption, for each stage of the supply chain i.e. growing, processing, packing or holding, and retailing	Identity of each supply chain actor / facility from: · Growers: USDA or state agricultural records · Processors and Distribution Centers/Warehouses: Required FDA registration information for · Retailers: Professional data collectors
Links and flows	Transition probability matrix ***F*** of normalized flow proportions f_{uv} out of node *u* in stage $n \in [1, N-1]$ and into adjacent node *v* in stage *n+1* during τ_P	Proportion of total volume of commodity sent from supply chain actor *u* in one stage to *v* in the subsequent stage, as a fraction of total volume handled by *u* over temporal period of analysis τ_P, e.g. week, growing season, year	Total volume of commodity handled by *u* sent to each trading partner *v* during τ_P, as recorded in traceability data or other transaction records (for *idealized* model), or estimated by producer /processor/packer/retailer (for *approximate* model)
Location of / distance between nodes	Matrix ***D*** of distances d_{uv} between adjacent node *u* in stage *n* and *v* in stage *n+1*, for all stages $n \in 1,...,N$	Shortest path road network distance between any traders *u* and *v* in all supply chain stages	Road network distances can be found for any two points on a map using GIS software. Geographical coordinates for each node found from: · Growers: USDA or state agricultural records · Processors and Distribution Centers/Warehouses: Required FDA registration information for · Retailers: Professional data collectors
Storage time delay for nodes	Storage delay distribution for commodity α at each supply chain node *u* in stages $n \in [2, N]$, with mean $\mu_{\alpha,u}$	Distribution of delay times, or the difference between receiving and dispatching times, associated with storage at each supply chain node	Sample of delays for commodity α at each supply chain node *u* during τ_P, as recorded in traceability data or other transaction records (for *idealized* model), or estimated by producer /processor/packer/retailer ((for *approximate* model)

Table 5.1. Data categories and sources to be used in creating the *idealized* and *approximate* network model for application to the foodborne disease detection problem.

5.1.2. Problems with idealized solution

The critical assumption underlying implementation of the idealized model is access to perfect data. There are two major problems arising from this assumption and they are polarizing: (i) acquiring and organizing information at this level of detail, and (ii) the compliance burden associated with succeeding.

Problem 1: Data Collection and Organization

We first discuss the feasibility of acquiring and organizing this data, and in particular, the transactional data necessary for representing link/flow and mean storage time delays that presents the greatest challenge owing to its abundance and the frequency with which it would need to be updated.

In principle, the necessary data does exist. Many food industry companies have an electronic system in place for recording transactions for proprietary business analysis, and more and more of these firms are adopting comprehensive technology-enabled traceability systems to capture the real-time movement of products through the supply chain (Storoy et al. 2013, McEntire and Bhatt 2013). If no official system has been set up, through due course of business, transactional data is generated in "native" forms such as bills of lading, purchase orders, harvest records, production records, shipping invoices, etc., or for tax and audit purposes (McEntire and Bhatt 2013).

The legal right to this data is another question. Trade relationships are proprietary information that businesses are often reluctant to share, since they constitute competitive information that presents advantages in the low-margin retail food industry. While each facility must participate with at least the "one-up, one-back" recordkeeping mandated by the Bioterrorism Act of 2002, this data is proprietary and must be shared with FDA only when responding to an incident. However it is possible that this obstacle might be overcome by new legislation passed under the landmark 2011 Food Safety Modernization Act (FSMA), the first major reform to food safety in 70 years, was signed into law. Among the Act's many provisions is an extension of traceability requirements. Specifically, FSMA extends the authority of the FDA to establish, as appropriate, "a product tracing system to *receive information that improves the capacity to effectively and rapidly track and trace food* that is in the United States or offered for import into the United States." The Act does not include specific provisions for implementing the law, and the regulatory timeline for determining the requirements is still developing. So far, the data gathering stage has begun and two product tracing pilots have been carried out for the purpose of exploring for how technology can be used by investigators to improve the traceback process. The FDA has yet to initiate the rulemaking process, and considerable work remains to be done before the agency will be in a position to decide upon an appropriate federal track and trace solution. Is it possible to imagine that access to the data necessary for the traceback methodology developed in this thesis could be made possible through this Act.

Assuming FSMA could expand the regulatory mandate to this data, its organization presents yet another challenge. The challenge is greatest for processed or otherwise multi-component foods containing dozens of ingredients. Even if many companies have robust traceability systems for efficient capture, storage, and communication, the necessary data will inevitably be recorded in different forms and levels of ready usability. Combining information from multiple platforms presents practical challenges, which are exacerbated for data recorded manually or in a native form unspecific to the task of traceability (e.g. invoice data). Ultimately, however, this is a surmountable problem faced by all organizations that deal with big, unwieldy data, and one that is being addressed at both the federal and industry levels. For example, as part FSMA, the FDA was required carry out product tracing pilots for the purpose of identifying and evaluating methods for how technology can be used by investigators to enhance the speed, effectiveness, and accuracy of the product tracing process. One of the main findings of the pilot study, conducted by the Institute of Food Technologists (IFT) under contract with the FDA, was that a lot of the key information necessary for traceability is already being captured in data systems, but that the way in which firms accessed and transmitted product tracing data varied widely. To connect these pieces of information into a package or system that the FDA can use, the IFT developed a specific set of requirements for improving capture and reporting of records, including standardized electronic mechanisms that would allow efficient aggregation and analysis of data (McEntire and Bhatt 2013). Movement towards a standardized recording and reporting system has emerged in industry as well. An organization consisting of growers, processors, retailers and foodservice companies has formed the Produce Traceability Initiative (PTI), whose purpose is prescribing a chain-wide standard format for electronic product traceability.

Problem 2: Compliance burden

In theory, a coherent framework for collecting and organizing the data could be implemented. On the other side of this, however, is the polarizing issue of the compliance burden. The more comprehensive the demands for data collection, the greater the burden on industry to comply and supply the information. Furthermore, the idealized system would require all members of the supply chain to participate.

A precedent for data collection and compliance has been set by the Bioterrorism Act of 2002, however the detail is many levels removed from what would be necessary. Firms under FDA's jurisdiction are required to officially register with the agency, providing their address, parent company name, trade names, and food product categories, among other requirements (H.R. 3488). After FSMA, the regulation was amended to require that facilities renew their registration every other year (H.R. 2751). This information in essence documents the location and supply chain stage identity of nodes in interior stages of the distribution network, e.g. processors, packers, and distribution centers. Growers, ranchers, and producers of raw commodity of any type are under the jurisdiction of the US Department of Agriculture (USDA) and thus are not included, however their registration information sits with the USDA's Farm Service Agency (FSA). Retailers too are not included, however their location information is available through professional data providers and marketing services such as the Red Book and the Produce Blue Book, which collect and market data under categories spanning the food and transportation industries (Red Book Credit Services 2016; Produce Blue Book 2016). Other requirements based on the Bioterrorism Act include having firms know who they received products from and to whom they were sent ("one up, one back" tracing); growers and retailers are still exempt. Furthermore, this data must be shared with FDA only when responding to a safety incident or when there is "reasonable cause" to make inquiries. While this information plays an essential rule during current traceback investigations, it is far from forming the idealized model, and the additional requirements necessary to get there would represent a significant increase in compliance burden on industry.

Furthermore, the degree that federal government should be able to regulate private business and trade and require the type of data necessary for the "idealized" solution is itself a very decisive topic. The appropriate strength of federal powers over private business has been a topic of extensive and often very passionate debate in this country since its inception and is very likely to indefinitely remain a major subject of political, economic and academic discourse. The Constitution provides objectively vague powers for federal regulation of business, most notably in the Commerce Clause (Article I, Section 8, Clause 3). The multiple interpretations of this clause have generated a significant branch of constitutional jurisprudence. More broadly, the differing policies on what the clause represents, the degree of separation between private enterprise and federal regulation, is a central distinguishing aspect of our two main political parties.

5.2. Alternative Network Solutions

We have presented two critical concerns regarding the feasibility of accessing, collecting, and organizing the idealized data. The significance of each issue and the fact that solving one exacerbates the other means there are significant challenges to implementing this modeling framework. While it can remain an ultimate goal, the FDA needs solutions to the traceback problem that are implementable in the short-term. In the following, we present three alternatives to the idealized model, each with its own strengths and weaknesses, but each requiring data at reduced granularity.

5.2.1. Approximate Modeling Framework

The first alternative is a modeling framework described by the same network characteristics and level of detail as the idealized model, but assembled from approximate rather than exact traceability data. Determining flow volumes and parameter values from the full data may be oversupplying the information necessary for the traceback methodology to identify *probable* outbreak origin locations. Instead of requiring documentation of the existence of every pair of trading supply chain actors and total magnitude of trade over each link, link/flow data could be assembled by requesting that each facility submit a list of the full set of business partners they (expect to) trade with, accompanied by an *estimate* of the proportion of total volume of production conducted with each actor. These estimates could be based on historical values, together with any forecasted changes in the upcoming time period. Similarly, storage time delay data could be collected through a less exhaustive process in which each facility is requested to submit an *estimate* of the average, minimum, and maximum distribution times in place of their full record of logistical data. Within some uncertainty, estimates should still provide adequate information to solve the probabilistic traceback problem. Data on the identity and location of the nodes would be collected by the same approach as in the idealized model, summarized in Table 5.1.

The clear benefit of this approach to data collection is that it would considerably reduce both the burden both on industry to supply the information, and on the central body responsible for organizing it. Indeed, the feasibility of implementing this approach, and doing so under conditions of voluntary compliance, has been demonstrated by a group of researchers working with Canada's Public Health Agency and Agriculture and Food Research and Development Centre. The goal of the project was to model the system-wide Canadian packaged lettuce retail supply chain. To simplify the task, the project focused on the five largest Canadian food retailers, which account for approximately 79% of food sales in Canada. A survey of growers, processors, distributors, and retail outlets was conducted to request information on (i) the volume of lettuce handled by each facility, (ii) the proportion supplied to each downstream node, and (iii) the number of days the product typically spends in each facility after production/arrival before being delivered to downstream nodes. From this information, the researchers were able to assemble a system-wide network model representing the probability of movement between actors in a supply chain.

This case study presents a first step towards demonstrating the implementability of the modeling approach for traceback purposes. The next steps would involve forming a *complete* picture of the supply chain including all businesses, and ultimately, validation for use in combination with a traceback methodology. The reliance on cooperating industry members to provide estimates introduces an uncertainty and a reduction in accuracy, both of which would be extremely difficult to quantify in practice. Furthermore, despite this demonstration that it can be implemented, the approach still presents similar drawbacks to the original model regarding feasibility, since significant data would need to be collected and updated on a regular basis to generate a database of networks for various food products. This constraint will limit the application of the methodology from tracing processed foods containing dozens of individual ingredients. One approach could be to focus on improving the accuracy of constructing networks for high priority, single-item foods, since simple foods cause the majority of the foodborne illnesses in the United States; in particular, 85% of those illnesses are attributed to produce (fruit and vegetables), poultry, dairy, eggs, and seafood alone (Painter et al. 2013). Even in the case of a single item, however, maintaining a comprehensive supply chain database would be onerous, even if the update cycle time τ_P is infrequent.

5.2.2. Compressed Modeling Framework

The second alternative we present is a *compressed* modeling framework in which only the first stage and last stage nodes are represented, that is, only the producer and the retailer stages. No knowledge about the identity or location, flows entering or leaving, or storage time delays at nodes in interior stages of the network is assumed. Instead, nodes represent supply chain actors engaged in growing and selling, but not processing, packing, or holding, food for consumption. Links in this condensed view represent the existence of at least one path between a producer and a retailer. Flows represent the proportion of total volume of commodity sent from supply chain actor u in the producing stage eventually each connecting retail node v along all paths traveled, as a fraction of total volume produced by u over temporal period of analysis τ_P.

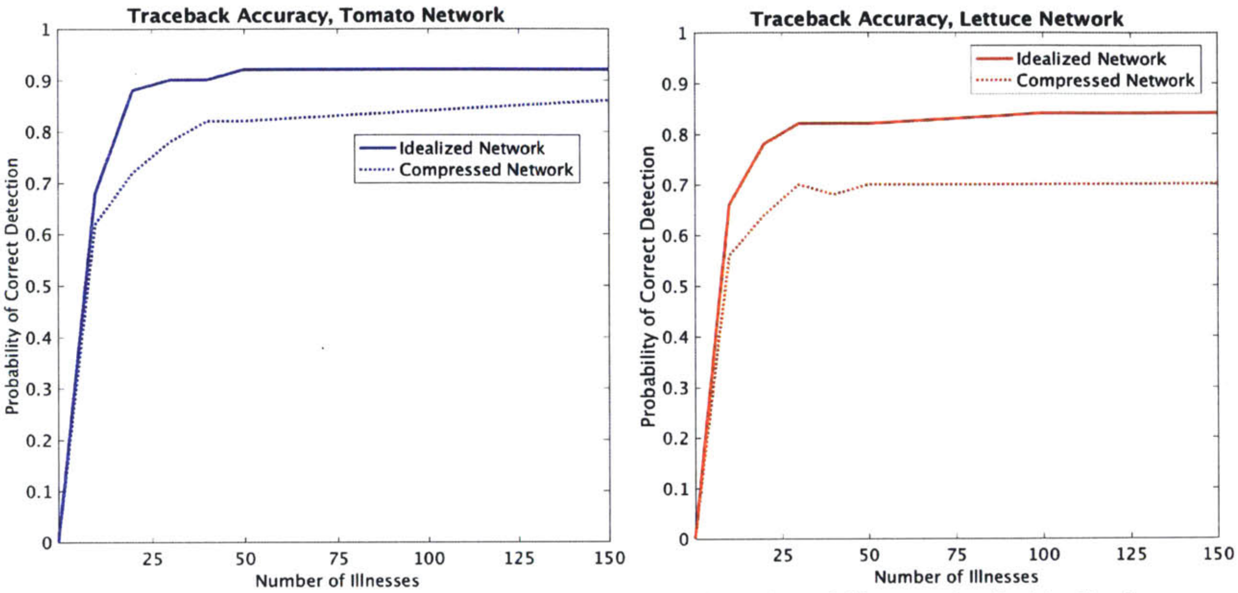
Implementing this modeling framework would represent a major decrease in the compliance burden, since the full set of *compressed* link/flow data could be formed by collected information at either only the first or last stage. Because it would be considerably more difficult to track the ultimate origin of products leaving a producer than to document the provenance of products arriving at the retailer, we recommend collecting data at the retail stage. Flow proportions could be calculated if total volume of commodity received at each retailer v from all points of origin u over the period τ_P were recorded. Since identifying point of origin does not represent "one-back" traceability and would not be available from "native" data records, to determine these volumes, it would be necessary to introduce additional track-and-trace requirements. The information would only need to document the business name or identification of the producer, which could be achieved by attaching some sort of labeling device at the producer level that would follow each commodity through its path to retail. This information is already attached to many individually packaged food items, or for bulk or loose items, the pallets or boxes these products are transported in. Of course, an in-depth analysis of the state of labeling in the industry would be necessary to determine how feasible it would be to achieve full compliance.

Traceback Model Input	**Definition**	**Description**	**Data To Collect**
Nodes	Supply chain nodes are actors u in stage 1 and v in stage N	Nodes represent supply chain actors engaged in growing and selling, but not processing, packing, or holding, food for consumption	Identity of each supply chain actor / facility from · Growers: USDA or state agricultural records · Retailers: Professional data collectors
Links and flows	Transition probability matrix F of normalized ***aggregate*** flow proportions f_{uv} from node u in stage $n=1$ eventually reaching descendent node v in stage $n=N$ from each ancestor node u in stage $n=1$, in aggregate over all possible paths of transmission, during τ_P	Proportion of total volume of commodity sent from supply chain actor u in growing/ production stage eventually reaching each connecting retail node v, as a fraction of total volume produced by u over temporal period of analysis τ_P, e.g. week, growing season, year	Total volume of the commodity received at each retailer v from all points of origin u over the period τ_P, as recorded in invoices or other transaction records, or estimated by producer /processor/packer/retailer

Table 5.2. Data categories and sources to be used in creating the *compressed* network model.

Since estimates of path lengths are not possible without the locations of interior stage nodes, applying this framework for outbreak traceback would mean applying an abridged version of the traceback methodology. Clearly, without the interior stages, it is not possible to determine the time delay density associated with individual paths traveled through the network. Accordingly, the entire spatio-temporal component of the algorithm must be disregarded. The resulting traceback algorithm determines the source PMF using only the aggregate probability flow from first stage to last stage nodes.

To evaluate the combined model-and-method approach theoretically, we applied it to the Tomato and Lettuce networks from the previous Chapters. The results are qualitatively illustrated in Figure 5.1, which plot the Traceback Accuracy as a function of number of illnesses for the idealized network and original method combination against the *compressed* network and abridged method combination for (a) the Tomato and (b) the Lettuce network. Table 5.3 quantifies the percentage decrease in accuracy at various intervals in number of reported illnesses. There is a clear loss in performance with the *compressed* network, though the drop is not extremely significant, ranging between 5 to 15% for both the Tomato and Lettuce networks.



Figures 5.1. Expected Traceback Accuracy as a function of number of illnesses for the idealized network and original method combination against the *compressed* network and abridged method combination, for **(a)** the Tomato and **(b)** Lettuce networks.

Traceback Accuracy Reduction With *Compressed* Modeling Framework		
Number of Illnesses	% Difference Tomato Network	% Difference Lettuce Network
10	-4%	-7%
20	-14%	-11%
30	-11%	-10%
40	-7%	-11%
50	-9%	-10%
100	-7%	-12%
150	-6%	-12%

Table 5.3. Percentage drop in traceback accuracy between the idealized network and original method combination against the *compressed* network and abridged method combination, at various intervals in number of illnesses, for the Tomato and Lettuce networks.

The drop in performance in the *compressed* case highlights the value of the spatio-temporal component to source localization, while the relatively minor reduction demonstrates the robustness of the aggregate probability component. The primary explanation for the small drop in performance is the uncertainty in the time delay density functions, which compounds over the combination of the multiple random variables representing the time a contaminated commodity takes to traverse a step along the path between contamination and eventual illness, each with its individual uncertainty. Still, traceback performance is clearly inferior without the temporal contribution and further analysis would be necessary to quantify the loss in benefits to public health and industry when combining this approach with the investigation strategies of Chapter 4.

There are clear advantages to implementing this network modeling framework over the idealized or *approximate* cases. With link/flow information needed to be collected only at the retail stage rather than for the producer, processor, and storage stages, the compliance burden shrinks considerably. Some additional traceability data would be necessary, but as discussed above, much of the data necessary to implement this approach is already fully available. Theoretically, traceback performance with the *compressed* approach is relatively robust to the loss in data on interior stages of the network; the drop is minor enough that it might be justified by the huge decrease in the compliance burden of implementing this approach. Still, the theoretical results assume perfect data, which reduced as it may be, would still require compliance by all members of the supply chain required by the model, i.e. retailers. Furthermore, the model could only be implemented after the necessary additional traceability requirements are determined, mandated, and the information collected. Ultimately, the major drawback of all of the methods presented thus far is the many steps removed the proposed approaches are from implementation. In all three cases, the specific data requirements and standards would need to be determined, their collection mandated by regulatory law, and a system for receiving and organizing the resulting data implemented by the FDA or other central authority responsible for data management.

5.2.3. Regionally Aggregated Modeling Framework

We now propose a third modeling framework alternative with a major advantage: a *regionally aggregated* network structure constructed using publicly available data sources without requiring data collection by a central authority. This alternative represents an *aggregated* modeling framework working at the regional or state level, where nodes are modeled as all actors of a supply chain type located in a specifically-defined regional area. For example, if working on a state level, all growers in the state of Massachusetts would be aggregated into a single Massachusetts "grower" node. The proposed framework is based on a methodology developed recently by researchers at Kühne Logistics University and the Technical University of Darmstadt in Germany. In this section we present an overview of the methodology as interpreted from the perspective of the modeling framework presented in this thesis; we will evaluate the appropriateness and accuracy of the methods for the source detection problem in future work (see Chapter 6). The overview of the data types, their interpretations in the foodborne disease problem, and the data sources to collect is summarized in Table 5.4. The data to be collected and methods for deriving the model categories nodes and links and flows come directly from the work presented in (Balster and Friedrich 2016). Storage time delay at supply chain nodes is not treated in their work; plausible suggestions for data sources are provided below, which will need to be verified through further research and discussion with the model's developers.

Traceback Model Input	Definition	Description	Data To Collect
Nodes	Supply chain nodes are regional districts $u_{r,n}$ representing operations of stage *n* in region u_r, for all stages $n \in 1,...,N$	Nodes represent the aggregate of all supply chain actors engaged in preparing food for consumption within a clearly delimited regional district. There are as many nodes for each district as there are supply chain stage types operating in that district, i.e. growing, processing, packing or holding, with the exception of retailing, which are modeled as one node per retailing company	Identity of each region in which activity of supply chain stage type *n* is conducted, determined from existing data sources – public authorities, food-related associations, and professional data providers
Links and flows	Transition probability matrix ***F*** of normalized flow proportions f_{uv} out of node $u_{r,n}$ in region u_r and stage $n \in [1, N-1]$	Proportion of total volume of commodity handled by actors of supply chain stage type *n* operating within region u_r, distributed to supply chain actors of subsequent stage operating within region v_r, as a fraction of total aggregate	Total volume of commodity handled by u_r sent to $v_{r,n+1}$ during τ_P, compiled from existing data sources – public authorities, food-related associations, and professional data providers

	and into adjacent node $v_{r,n+1}$ in region v_r and stage $n+1$, during τ_p	volume handled by actors of stage *n* within district u_r over temporal period of analysis τ_p, e.g. week, growing season, year	
Location of / distance between nodes	Matrix ***D*** of distances: · d_{uv}, for the average distance between distinct regions u_r and v_r; · d_{uu}, for the average distance between two points within the region u_r	· For transport from one stage to the next between distinct regions u_r and v_r, the geographical distance between the spatial center of the two regions · For transport from one stage to the next within the same district u_r, the average distance between two points within the region	Geographical dimensions and coordinates for midpoint of each region, from spatial / GIS data
Average transport speed	Fixed parameter v_{avg}	Average velocity of food transport travel, including delays, within given country context	In the US, Average food retail truck transport speed collected by Bureau of Transportation Statistics (BTS)
Storage time delay for nodes	Storage delay distribution for commodity α at each supply chain node *u* in stages $n \in [2,N]$, with mean $\mu_{\alpha,u}$	Distribution of delay times, or the difference between receiving and dispatching times, associated with storage at each supply chain node	Sample of delays for commodity α at each supply chain node *u* during τ_p, as recorded in traceability data or other transaction records, or estimated by producer /processor/packer/retailer

Table 5.4. Data categories and sources to be used in creating the *regionally aggregated* network model. The modeling framework is based on the methodology presented in (Balster and Friedrich 2016) with the exception of the categories **location of / distance between nodes, average transport speed**, and **mean of storage time delay for nodes**; these sources are not directly referenced in (Balster and Friedrich 2016) and the suggestions here will need to be verified through further research and discussion with the model's developers.

The work presented in Friedrich (2010) and Balster and Friedrich (2016) define a methodology for modeling dynamic commodity flows for the entire food supply chain using existing data sources coming from public authorities, food-related associations, and professional data providers. The framework models the entire supply structure for a given commodity, differentiating between node stage types e.g. production, storage, and retail, within different *markets* operating within the supply chain: *wholesaling*, or sales to industrial consumers such as restaurants, hotels, schools, and prisons; *retailing*, sales directly to consumer; *direct sales*, or sales purchased by the consumer directly from the commodity's producer, e.g. farmers markets; and diversion of commodity into the production of a compound food type. The inclusion of this final category means that the framework models multiple foods simultaneously, including combinations and transformations of food products, making it fully extendable to tracing processed foods containing dozens of individual ingredients.

For the reasons discussed in this Chapter, readily-accessible supply chain data is only available on an aggregate, regional level. The same resolution is therefore adopted in the model: nodes $u_{r,n}$ represent the aggregate of all supply chain actors / facilities engaged in preparing food for consumption in stage *n* within a clearly delimited regional district u_r. Multiple stages exist for each node: there are as many nodes for each district as there are supply chain stage types operating in that district, i.e. growing, processing, packing or holding, with the exception of retailing companies, which are modeled independently, i.e. one node per retailing company. Links and flow proportions thus represent the proportion of total volume of commodity handled by actors of one supply chain stage operating within a given region, distributed to supply chain actors of the subsequent stage operating within the same or another region, over the temporal period of analysis τ_P.

First, annual flows are determined from the publicly available data sources using a trade flow Gravity Modelling approach. A detailed explanation of the methodology can be found in Balster and Friedrich (2016), which we overview here. The Gravity Model is a modified version of Newton's gravitation model, assuming that the probability of two market actors trading with each other is proportional to the supply and demand of the respective actors and indirectly proportional to their distance to each other (Anderson and Van Wincoop 2003). The Gravity Model is calibrated using transport data collected annually by the Federal Transport Plan in Germany and the Bureau of Transportation Statistics (BTS) in the US. The annual flows are calibrated using federal transport data following a procedure suggested in (Balster and Freidrich 2016), which extends methods from the class of multi-regional input-output models (Cascetta 2008). After determining annual flows, the model calculates inventories for each group of actors and every region. The inventories are recalculated incrementally every day, considering the production, relocation of food products, and consumption. The result is a comprehensive analysis of day-to-day inventories in and flows among actors and regions. In application to the foodborne disease traceback problem, thèse flows would be aggregated over the temporal period of analysis τ_P.

In the *regionally aggregated* view, the data category *mean of storage time delay at nodes* is represented by the mean of the delay times averaged over all supply chain nodes of type *n* operating within a region u_r. Since the characteristic delay distribution representing all nodes of stage u_r should not be distributed differently than that aggregating over all nodes in another region v_r, one mean storage delay term can be used for each supply chain stage *n*.

To determine the matrix ***D*** of distances in the *regionally aggregated* model, two terms must be accounted for: (1) transport from one stage to the next between distinct regions u_r and v_r, and (2) transport from one stage to the next within the same district u_r. For (1), the most direct proxy would be to use the geographical distance between the spatial center of the two regions. For (2), we suggest determining the average distance between two points within a shape of the same dimensions as u_r.

So far, the researchers at Kühne and Darmstadt have illustrated this methodology in application to the German food supply system for the year 2012. Their implementation works on an aggregate level of 402 regions within Germany as well as the 50 most important trading nations and includes 51 commodity groups. In Figure 5.2, we present a representation of the aggregated supply chain model system applied to single-item food type A. This example includes four supply chain stages: production, storage/holding/packing, sale at retail stores, and consumer purchasing. The different colors in the figure represent the four types of *markets* operating within the supply chain: *wholesaling*, or sales to industrial consumers such as restaurants, hotels, schools, and prisons; *retailing*, sales directly to consumer; *direct sales*, or sales purchased by the consumer directly from the commodity's producer, e.g. farmers markets; and diversion of commodity into the production of a compound food type. Each market will operate independently until combining at the consumer purchasing level. More information on this food system model is provided in Friedrich (2010) and Friedrich and Balster (2016). Flows across stages are permitted in this network view, as can be seen by links in the direct sales and wholesaling market streams. These across-stage links can be accommodated by the source traceback framework without any extensions to the methodology.

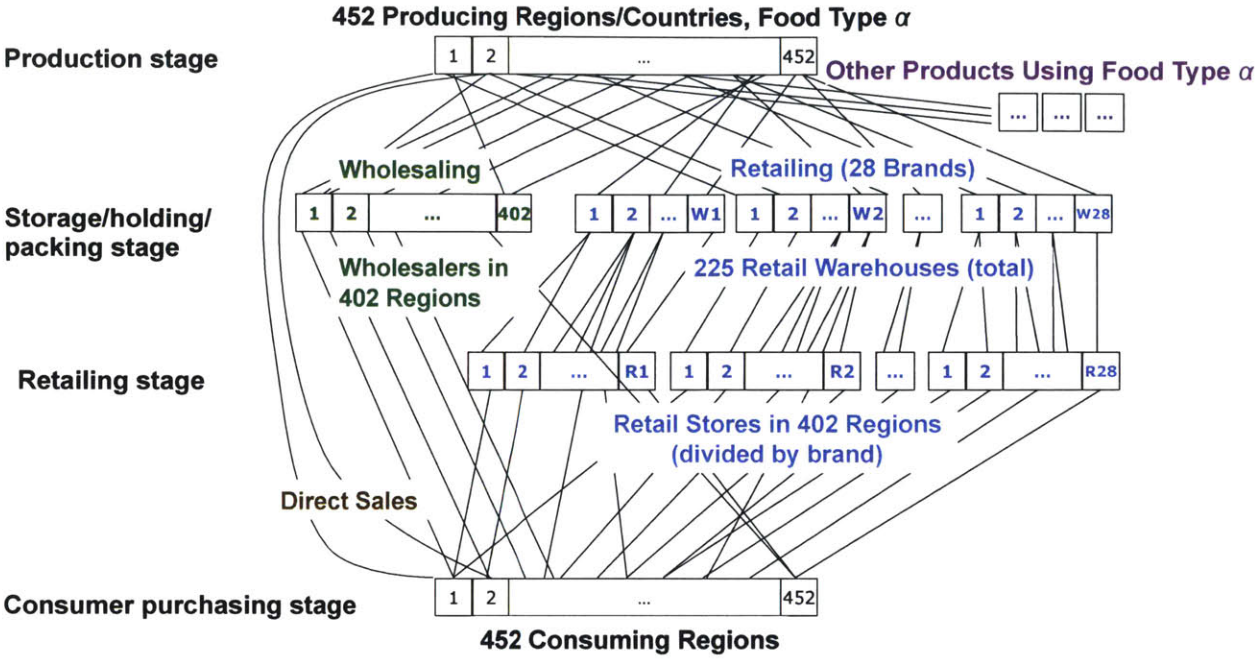

452 Producing Regions/Countries, Food Type α
Production stage
1
2
...
452
Other Products Using Food Type α
...
Wholesaling
Retailing (28 Brands)
Storage/holding/ packing stage
402
W1
W2
W28
Wholesalers in 402 Regions
225 Retail Warehouses (total)
Retailing stage
R1
R2
R28
Retail Stores in 402 Regions (divided by brand)
Direct Sales
Consumer purchasing stage
452 Consuming Regions

Figure 5.2. Representation of the *regionally aggregated* network modeling framework applied to single commodity type α . Four supply chain stages are represented: production, storage/holding/packing, sale at retail stores, and consumer purchasing. The different colors in the represent the four types of *markets* operating within the supply chain: *wholesaling*, *retailing*, *direct sales*, and diversion of commodity into the production of other products using food type α .

There are significant advantages of employing a modeling framework based on existing sources and without requiring additional data collection. In particular, there would be zero data collection burden to industry, and the model is immediately implementable and ready to use for traceback purposes. Furthermore, because the data is readily available and electronically accessible, networks could be generated more or less instantaneously at the time of an event, for single and multiple-ingredient foods. For all other approaches presented in this Chapter, a bank of network models would need to be constructed in advance and updated at regular intervals τ_P .

5.3. Recommended network modeling approach

On the basis of its ready implementability, extendibility to multiple food types, and potential for accuracy within regional limits, we recommend deploying our traceback methodology with the KLU food distribution network model. In an upcoming research collaboration, we will work together with researchers at the KLU to integrate their network model with our traceback methodology to form a *holistic traceback framework*. Using this combined model and method approach, we will seek to demonstrate the ability to identify the origin of recent outbreaks that have occurred in Germany. This work is already planned, as described in the Conclusion (Chapter 6). Success in correctly identifying the source of these outbreaks will be an important step in validating the accuracy and effectiveness of our methodology for identifying the source of large-scale outbreaks of foodborne disease.

This work will seek to (i) verify the feasibility of combining the dynamic commodity flow model with the spatio-temporal method for source localization developed in this thesis, and (ii) determine the loss of granularity in localizing the source resulting from the aggregation of nodes to regions and thus the inability to differentiate between feasible sources within a region. Clearly, the accuracy will be directly limited by the granularity of the regional unit of analysis and the distribution of production across the map, with boundaries that mean that producers are not all clustered into a few regions. For example, while the Lettuce network modeled in this thesis represents only three growing regions, each of these regions is subdivided into counties, whose borders we have not included in the model.

5.4. Conclusions

Implementing the traceback methodology developed in this thesis relies on access to a database of representative network models for various food types. However, acquiring and organizing food supply system data presents multiple practical challenges. First, food distribution networks are markets characterized by inherent stochasticity. While many trade relationships are enduring, others may be based on transitory spot-markets that are difficult to know and to model. Second is the legal right to the data, though this obstacle might be overcome by new legislation passed under the 2011 Food Safety Modernization Act (FSMA). Trade relationships are proprietary information that businesses are often reluctant to share, since they constitute competitive information that presents advantages in the low-margin retail food industry. While each facility must participate with at least the "one-up, one-back" recordkeeping mandated by the Bioterrorism Act of 2002, this data is proprietary and must be shared with FDA only when responding to an incident. Third is that even if this information is made available to regulators, collecting and organizing the data in a well-maintained system would present extensive practical data-management challenges. Foods of today are complex and outbreaks can occur in foods containing dozens of ingredients.

Clearly, the performance of the combined model-and-method approach will ultimately depend on the properties of the underlying network, including the granularity of the model and the representativeness of the data informing the model. Due to the challenges associated with collecting the data, it would be opportune to collect the minimal data necessary to achieve high traceback performance without oversupplying it. In this chapter, we seek to understand the parameters of the "sufficient" network data. We propose multiple approaches to modeling the supply chain network, each at a different level of detail or granularity. For each approach, we suggest a means to collect the necessary data and then examine (i) the feasibility of collecting this data and (ii) the potential traceback accuracy achieved when implementing the model together with the traceback method. On the basis of this analysis, we conclude that a network model based on a recently developed food supply modeling methodology developed by researchers at Kühne Logistics University (KLU) and the Technical University of Darmstadt in Germany demonstrates the strongest potential for overcoming the said challenges. This model is both readily implementable, utilizing only existing, readily available data sources coming from public authorities, food-related associations, and professional data providers. It is also comprehensive, covering the supply of 50 different foods and their interactions, making it extendable to tracing processed foods containing dozens of individual ingredients. The next step will be to deploy our traceback methodology with the KLU food distribution network model, integrating the two to form a *holistic traceback framework*. Next steps in the evaluation and implementation of this framework are presented in Chapter 6 (Conclusions).

Chapter 6: Conclusions

6.1 Overview

This thesis is focused on the increasingly dangerous, but largely understudied problem of large scale outbreaks of food-borne disease. The overarching goal of this thesis was to develop techniques to efficiently identify the source of a large-scale food-borne disease outbreak while contamination-caused illnesses are still occurring and provide food-safety practitioners with a framework for making decisions based on source identification, thereby resolving investigations earlier and averting potential illnesses. To meet this goal, we developed a holistic traceback framework centered on a network-theoretical approach for rapid identification of the source of foodborne contamination events. The primary contribution of the network-theoretical approach is a Spatio-Temporal Traceback (STT) algorithm, which uses backward induction and network analysis to determine the probability that any location in a network is the outbreak source.

To deploy the methodology in the event of an outbreak, the first step would be to generate a supply chain network model for the food or foods in question according to the *regionally aggregated* modeling framework. Second, given the location and timing of the reported cases of foodborne illness, the source detection framework would be applied to identify high probability sources of contamination. Finally, strategic recommendations regarding allocation of investigative resources for search effort and mitigation measures would be determined by applying the intervention mechanisms described in Chapter 4. All stages could be implemented more or less instantaneously, with the data sources and analytical models all being computer implemented.

The traceback framework was subjected to an extensive study to evaluate the detection performance and robustness across multiple outbreak scenarios and network structures. In addition, studies were conducted to measure the benefits of the methods in comparison to existing approaches: a practical heuristic meant to model current methods applied in practice, and a best-in-class method presented in the literature. The results from simulation studies across a range of realistic outbreak scenarios and network structures demonstrate the methodology is highly accurate and efficient, consistently outperforming existing approaches by a significant margin. Theoretically, these results demonstrate the suitability of our specific methodological approach to the problem of localizing the source of foodborne disease outbreaks. Practically, they suggest that if implemented by investigators, our method can contribute to the traceback investigation process in significant ways. Investigations previously unapproachable could be successfully and conclusively resolved. More importantly, for those cases when identification is possible early enough in an outbreak's progression and quick action is taken, a substantial fraction of the illnesses might be averted. These conclusions will ultimately need to be validated beyond the simulation results presented here.

To perform the evaluation studies, a probabilistic simulation network model was developed, primarily because there was no prior existing network model that allowed for realistic simulations across multiple food types and outbreak scenarios. The model was structured as a directed network with four stages of supply and distribution, representing the flow of US's complex food-system network. The simulation model allowed us to test algorithm by generating contamination events and observing the algorithm's ability to trace reported outbreaks back to a set of possible sources.

Once the accuracy and robustness of the STT algorithm was demonstrated, we developed strategies for real-time investigation response based on the traceback inference methodology. These strategies include determining (i) when and to which potential source candidates to send out investigators to sample predictions and (ii) when to message the public about the outbreak source and how to frame the statement. The combination of the network-based inference algorithm with the decision-making tool formed a holistic traceback framework.

Lastly, while the potential benefits of the holistic framework are promising, it has heretofore only been tested on simulated food systems networks with probabilistic outbreak scenarios. Additionally, it requires a dynamic data network of food production, processing, distribution and retail that is currently infeasible or impossible for a food safety regulatory agency to acquire or maintain. However, there is strong promise that a *regionally aggregated* network model, as described at the end of Chapter 5 can be used to overcome this latter shortcoming of the traceback framework. Testing of the STT algorithm on a *regionally aggregated* network model using real food distribution and historical outbreak data will be immediate next steps in the development of the work presented in the is thesis.

6.2 Key Findings

This thesis has contributed a novel approach and framework to outbreak traceback investigations: a computer-based methodology that has the potential to efficiently identify the location source of large scale, distributed outbreaks of foodborne illness with high accuracy. The following are key findings and contributions from the development and testing of this network based traceback framework.

Advantages of the proposed traceback methodology in comparison with current approaches

- Immediately identify *feasible* source locations
- Rank *feasible* sources by likelihood each one is *true source*
- Use ordered ranking to create systematic investigation strategies
- Low financial and opportunity cost to implement
    - Low financial cost since the traceback system would function on a computer model at very low cost to implement and no cost to operate.
    - Low opportunity cost because generation of investigation recommendations can be done at no exclusion of other approaches
- Leverage all data available:
    - All case reports, including initial or "tentative" cases; Comprehensive system network data

Improved accuracy and efficiency in source identification during large scale outbreaks

While the specific benefits will be determined by the particular outbreak scenario, general conclusions can be drawn regarding the potential impact of the mitigation-based approach to developing and implementing interventions. In particular, the computational results presented here suggest that benefits can be measured on three important dimensions:

1. ***Methodology can outperform existing methods in traceback investigation by significant margin***, requiring a fraction of resources necessary to achieve the same level of benefits; in simulation results, the source is identified within a set of 5 top ranked candidates with 95% accuracy vs. 50 – 150 sources required by current methods
2. ***Many more investigations successfully and conclusively resolved:*** In simulation testing across a variety of distribution network structures, we found that the true source can be identified or narrowed down to a small, bounded set of possibilities with very high accuracy

and efficiency: the actual outbreak source was robustly ranked within the top 5% (1%) of feasible locations after 10% (30%) of the cases had been reported.

3. ***Identification is possible early enough in an outbreak's progression that a significant fraction of illnesses can be averted***; results suggest numbers between 40 – 60% of the illnesses that would ultimately result from simulated outbreaks

Broader understanding about the problem of source traceback in complex systems

- Characteristics of network topology have been identified to that improve / decrease detection performance; in particular, the degree of heterogeneity and connectivity in a network.
- Taken together, these insights provide a first step in understanding how the accuracy of detection depends on the structure of a network and on the stochastic evolution of the disease trajectory.

6.3 Future Work

As stated in the body of this thesis, these findings are derived from simulation results; live use of these techniques has yet to occur and may demonstrate features of the real problem inadvertently omitted from the modeling. To validate the relevance of the traceback system in times of real public health emergency, it will be necessary to demonstrate its ability to accurately identify outbreak origins and recommend effective investigation strategies. Extensive testing will be necessary to determine the utility to public health, measured in terms of how much earlier an investigation can be resolved and how many illnesses averted as a result.

6.3.1 Validation through application to historical outbreaks

The first step towards validation will be to integrate the holistic traceback framework with the *Aggregated Network Modeling Framework* described in Chapter 5, then demonstrate the ability to correctly localize the origin of historical outbreaks. To conduct this first step, we will work with the researchers at Kühne Logistics University and the Technical University of Darmstadt in Germany who developed the *Aggregated Network Modeling Framework* (planned during October – December 2016).

Basing the methodology on the *regionally aggregated network* is preferable for validation because the network is generated from publicly-available data. Because the *regionally aggregated network* was originally developed in Germany, we will verify the feasibility of combining the approaches and to demonstrate the ability to correctly localize the origin of historical outbreaks of foodborne disease in Germany. An example in practice would be the E. coli outbreak of 2011 linked to sprouts grown in district Uelzen, which lead to nearly 4,000 reported illnesses and 53 deaths (WHO 2011). Another possibility is the much smaller outbreak in 2014 of Salmonella Enteritis PT14b, traced to an egg producer in Bavaria (Bayern Ei), which led to 24 cases of illness in Germany (ECDC-EFSA 2014). We intend to apply the techniques to other significant outbreaks of foodborne disease in Germany occurring after 2010.

For each food, we will develop a distribution network model for the aggregated flows of the contaminated commodity (e.g. for sprouts and eggs within Germany). We will collect the case report data by location and date of illness onset from the Robert Koch Institute (RKI). We will then develop and implement strategies to tactically respond to the investigation at multiple time intervals in the course of the outbreaks' progression. This study will allow us to directly quantify the benefits of our tool through the comparison to the existing methods used in outbreak investigations. We well determine the effectiveness of the strategy developed for public service messaging through comparison to the measures taken during the actual investigations. Specifically, we will quantify how much earlier our approach would have been able to converge on the true outbreak source location than the historical investigation, and how many illnesses could have been avoided had a public service announcement been made at the time of detection.

To further extend this analysis, we will subject the combined *aggregated*-model-and-method system to extensive robustness tests to determine the accuracy when applied to various differing commodity types. We will make comparisons between commodities that exhibit different distribution patterns, for example choosing fresh seafood, which exhibits regionally concentrated production and thus cross-national distribution, and eggs, with production distributed throughout the country and thus highly localized distribution. We will compare the results from the case studies to the sensitivity analyses from synthetic outbreaks to draw conclusions about the dependence and limits of accuracy on the distribution structure of various food types. The successful completion of the proposed initial model-and-method validation Germany will provide useful information to food-safety regulators internationally. It will demonstrate that it can be applied in any region (defined by national boundaries or not) in which a food-safety regulator (or consortium of regulators) can benefit from rapid identifying the source of foodborne disease, as long as the statistical data required to model the network structures are available.

Ultimately, the methodology will need to be tested through live experiments in differing outbreak cases and country contexts: the US, the broader EU, and developing countries. However even then, in all application to historical cases the data available will be better than what would be available at the time of an outbreak, due to the delays and inaccuracies in case reporting. Real-time application of the tool during outbreak emergencies will ultimately be necessary to determine the utility to public health in terms of how much earlier an investigation can be resolved and how many illnesses averted as a result.

6.3.2 Extensions of the foodborne disease source identification problem

Future work should seek to combine the source identification methodology with other computational/ "big-data" methods for outbreak detection (Digital Disease Detection) and identifying the specific product source, creating a comprehensive system for outbreak response. This would include:

- Detecting the outbreak using Digital Disease Detection:
- Identifying food source using sales data
- Identifying location source: methods of this thesis

Future work should seek to expand upon insights derived from learning about dependence of traceback accuracy on network structure. This can inform the design of network structure that more robust to both propagation of contamination, as well as to facilitate traceback. For example, we are interested in evaluating the structural property of consolidation in the food supply chain, which is defined as the organization of production into fewer but larger plants or farms. It will be important to identify implications for the safety of the food supply, and for the proactive design of supply networks that limit propagation and improve tracebacks.

This thesis presents a novel approach to source identification of spreading agents in networks of food distribution. This may be the start for further research projects in modeling and source identification of spreading agents in complex networks. The fundamental requirement our methodology is access to a (mostly) known network structure characterized by weighted links and (mostly) known temporal dynamics. Given this data, the traceback methodology developed here can be adapted to other problem contexts that are growing in importance as our society increases in connectedness, such as

- Infectious disease + weighted, known transport network + case reporting times
- Hospital contagion + patient-provider networks
- Violence (or terrorism) + online social networks

References

Agricultural Marketing Service, United States Department of Agriculture (AMS, USDA). (2015). Market News Custom Reports. Accessed August, 2015. Available at: https://www.marketnews.usda.gov/mnp/fv-report-config-step1?type=movement

Ahumada, O. and Villalobos, J.R. (2009). Application of planning models in the agri-food supply chain: A review. European Journal of Operational Research 196(1):1-20.

Anderson, J.E. and Van Wincoop, E. (2003). Gravity with gravitas: a solution to the border puzzle. The American Economic Review 93: 170-192.

Anderson, M., Jaykus, L. A., Beaulieu, S., and Dennis, S. (2011). Pathogen-produce pair attribution risk ranking tool to prioritize fresh produce commodity and pathogen combinations for further evaluation (P3ARRT). Food Control, 22(12), 1865-1872.

Balster, A., and Friedrich, H. (2016). Dynamic freight flow modeling for risk evaluation in food supply. World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016.

Batz, M.B., et al. (2005). Attributing illness to food. Emerg Infect Dis 11(7):993-999.

Batz, M.B., Hoffmann, S., Krupnick, A., Morris, G., Sherman, D., Taylor, M., et al. (2004). Identifying the most significant microbiological foodborne hazards to public health: A new risk ranking model. Food Safety Research Consortium. Discussion Paper Series, Number 1.

Batz, M.B., Hoffmann, S.A., and Morris, J.G. Jr. (2012). Ranking the Disease Burden of 14 Pathogens in Food Sources in the United States Using Attribution Data from Outbreak Investigations and Expert Elicitation. Journal of Food Protection. 75(7): 1278-1291.

Beni, L.H., Villeneuve, S., LeBlanc, D.I., and Delaquis, P. (2011) A GIS-based Approach in Support of an Assessment of Food Safety Risks. Transactions in GIS 15(s1):95-108.

Bertolini, M., Bevilacqua, M., and Massini, R. (2006). FMECA approach to product traceability in the food industry. Food Control 17 (2), 137–145.

Beuchat, L.R. 1996. Pathogenic microorganisms associated with fresh produce. J. Food Prot. 59:204-216.

Blue Book Services (2016). Credit Ranking and Marketing Information for the Produce Industry. Accessed July, 2016. Available at https://www.producebluebook.com

Brockmann, D., and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. Science, 342(6164), 1337-1342.

Brockmann, D., David, V., and Gallardo, A. M. (2009). Human mobility and spatial disease dynamics. Reviews of nonlinear dynamics and complexity, 2, 1-24.

Brown, D.P. (2013). "Tropical Cyclone Report: Hurricane Barbara." National Hurricane Center (National Oceanic and Atmospheric Administration). Published August 19, 2013. Retrieved December 1, 2014.

Business Insights (2010). Food Safety and Traceability Strategies: Key hazards, risks and technological developments.

California Department of Public Health (CDPH) (2007). Investigation of an Escherichia coli O157:H7 Outbreak Associated with Dole Pre-Packaged Spinach, Final Report prepared by the California Food Emergency Response Team. March 21, 2007. http://www.cdph.ca.gov.

Cascetta, E., Marzano, V., and Papola, A. (2008). Multi-regional input-output models for freight demand simulation at a national level. Recent developments in transport modelling: Lessons for the freight sector, 93-116.

Centers for Disease Control and Prevention (CDC) (2006). Ongoing Multi-State Outbreak of Escherichia coli serotype O157:H7 Infections Associated with Consumption of Fresh Spinach. Morbidity and Mortality Weekly Report, 55(Dispatch); 1-2. September 26, 2006. http://www.cdc.gov/mmwr/preview/mmwrhtml/mm55d926a1.htm

Centers for Disease Control and Prevention (CDC) (2011). Multistate Outbreak of Listeriosis Linked to Whole Cantaloupes from Jensen Farms, Colorado.) December 8, 2011. http://www .cdc.gov/listeria/outbreaks/cantaloupes-jensen-farms/120811/index.html

Centers for Disease Control and Prevention (CDC) (2013). Outbreaks of Cyclosporiasis – United States, June – August 2013. MMWR. Morbidity and mortality weekly report, 62(43), 862.

Centers for Disease Control and Prevention (CDC) (2013b). Cyclosporiasis Outbreak Investigations — United States, 2013 (Final Update). December 2, 2013. Available at http://www.cdc.gov/parasites/cyclosporiasis/outbreaks/investigation-2013.html Accessed July, 2015.

Centers for Disease Control and Prevention (CDC) (2015a). PulseNet: The First Step in Identifying a Foodborne Outbreak. Accessed July 1, 2014. http://www.cdc.gov/pulsenet/outbreak-detection/index.html

Centers for Disease Control and Prevention (CDC) (2015b). Timeline for Reporting Cases of E. coli O157 Infection. Accessed July 1, 2015. http://www.cdc.gov/ecoli/reportingtimeline.htm

Centers for Disease Control and Prevention (CDC) (2015c). Timeline for Reporting Cases of Salmonella Infection. Accessed July 1, 2015. http://www.cdc.gov/salmonella/outbreaks/reporting_timeline.html

Centers for Disease Control and Prevention (CDC) (2015d). Multistate Outbreaks: Detecting a Possible Outbreak. Accessed July 1, 2015. http://www.cdc.gov/foodsafety/outbreaks/investigating-outbreaks/investigations/detection.html

Centers for Disease Control and Prevention (CDC). (2015e). Foodborne outbreak online database (FOOD).

Comin, CH, and da Fontoura Costa L. (2011). Identifying the starting point of a spreading process in complex networks. Phys Rev E, 84(5):056105-1--11.

Conrad, S.H., Beyeler, W.E., Brown, T.J. (2012). The Value of Using Stochastic Mapping of Food Distribution Networks for Understanding Risks and Tracing Contaminant Pathways. Int. J. Critical Infrastructures, Vol. 8, Nos. 2/3, September 2012, 216-224.

Decker, Kelly (2014). Preventing Agroterrorism. Interview by Mollie Halpern. Audio blog post. FBI,

DeWaal, C. S., Roberts, C., & Catella, C. (2015). Outbreak Alert! 1990-2011. Center for Science in the

Public Interest, Washington, DC.

Dobbie, J. M. (1968). A survey of search theory. Operations Research, 16(3), 525-537.

Doerr D, Hu K, Renly S, Edlund S, Davis M, et al. (2012) Accelerating investigation of food-borne disease outbreaks using pro-active geospatial modeling of food supply chains. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Use of GIS in Public Health. ACM: Redondo Beach, California: 44–47. doi 10.1145/2452516.2452525.

Dupuy, C., Botta-Genoulaz, V., Guinet, A. (2005). Batch dispersion model to optimise traceability in food industry. Journal of Food Engineering 70, 333–339.

The European Centre for Disease Prevention and Control and the European Food Safety Authority (ECDC-EFSA) (2014). Multi-country outbreak of Salmonella Enteritidis infections associated with consumption of eggs from Germany. The Joint ECDC–EFSA Rapid Outbreak Assessment Team. http://ecdc.europa.eu/en/publications/Publications/salmonella-enteritidis-rapid-outbreak-assessment-270814.pdf Accessed June 28, 2016.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian?". Bayesian analysis, 1(1), 1-40.

Fioriti, V., and Chinnici, M. (2012). Predicting the sources of an outbreak with a spectral technique. arXiv preprint arXiv:1211.2333.

Food and Agriculture Organization of the United Nations and World Health Organization (FAO-WHO). (2008). Microbiological hazards in fresh fruits and vegetables. Microbiological Risk Assessment Series. Meeting Report (prepublication version). Accessed July 2015. Available at: http://www.who.int/foodsafety/publications/micro/MRA_FruitVeges.pdf

Food and Drug Administration (FDA) (2001). Guide to traceback of fresh fruits and vegetables implicated in epidemiological investigations. Rockville (MD): The Division of Emergency and Investigational Operations, Office of Regional Operations, Office of Regulatory Affairs, FDA.

Food and Drug Administration (FDA) (2013a). Food Defense. FDA Releases New Tool to Help Prevent

Food and Drug Administration (FDA) (2013c). FDA Investigates Multistate Outbreak of Cyclosporiasis. November 21, 2013. Available at http://www.fda.gov/Food/RecallsOutbreaksEmergencies/Outbreaks/ucm361637.htm Accessed July, 2015.

Food and Drug Administration (FDA) (2014). What You Should Know About Government Response to Foodborne Illness Outbreaks. April 7, 2014. http://www.fda.gov/Food/ResourcesForYou/Consumers/ucm180323.htm. Accessed March 15, 2015.

Food and Drug Administration (FDA). (2013b). Bad bug book: handbook of foodborne pathogenic microorganisms and natural toxins. Accessed July, 2015. Available at: http://www.fda.gov/downloads/Food/FoodborneIllnessContaminants/UCM297627.pdf

Food Safety Modernization Act (FSMA) (2011). H.R. 2751 FDA Food Safety Modernization Act.

Friedrich, H. (2010). Simulation of logistics in food retailing for freight transportation analysis. Doctoral dissertation, Karlsruhe Institute for Technology.

Fritz, M., Schiefer, G. (2009). Tracking, tracing, and business process interests in food commodities: a multi-level decision complexity. International Journal of Production Economics 117, 317–329.

Golan, E., B. Krissoff, F. Kuchler, L. Calvin, K. Nelson, and G. Price. (2004). Traceability in the U.S. food supply: Economic theory and industry studies. Agricultural Economic Report No. 830, United States Department of Agriculture, Economic Research Service. Washington, DC.

Gonzalez, Marta C., Cesar A. Hidalgo, and Albert-Laszlo Barabasi. (2008). Understanding individual human mobility patterns. *Nature* 453.7196 (2008): 779-782.

Grady, Daniel, Christian Thiemann, and Dirk Brockmann. (2012). Robust classification of salient links in complex networks. Nature communications 3: 864.

H.R. 3448. (2002). Public Health Security and Bioterrorism Preparedness and Response Act of 2002 (2002). http://grants.nih.gov/grants/policy/select_agent/HR3448_Public_Health.pdf

Harlander, S. and Sholl, J. (2007). Software Systems for Food Safety and Defense. AIB International, May/June 2007: 9-10.

Harris, Jenine K., et al. (2014). Health department use of social media to identify foodborne illness-Chicago, Illinois, 2013-2014. MMWR Morb Mortal Wkly Rep 63.32: 681-685.

Harrison, Cassandra, et al. (2014). Using online reviews by restaurant patrons to identify unreported cases of foodborne illness—New York City, 2012–2013. *MMWR* 63.20: 441-445.

Hashemi-Beni, L., Otten, A., Fazil, A. McKellar, R., and Delaquis, P. (2015). A national produce supply chain database for food safety risk analysis. Journal of Food Engineering 147 (2015): 24-38.

Hashemi, B.L., et al. (2012). Spatio-temporal assessment of food safety risks in Canadian food distribution systems using GIS. Spatial and Spatio-temporal Epidemiology.

Institute of Medicine (2010). Enhancing Food Safety: The Role of the Food and Drug Administration. National Academies Press.

Intentional Food Contamination. FDA News Release. FDA, 13 May 2013. Web. 12 Mar. 2015.

James M. MacDonald, Penni Korb, and Robert Hoppe, USDA, Economic Research Service. (2013). Farm Size and the Organization of U.S. Crop Farming. Economic Research Report No. (ERR-152).

Kaufman, J., Lessler, J., Harry, A., Edlund, S., Hu, K., Douglas, J., Thoens, C., Appel, B., Käsbohrer, A. and Filter, M., (2014). A likelihood-based approach to identifying contaminated food products using sales data: performance and challenges. *PLoS Comput Biol*, *10*(7), p.e1003692.

Keeling, M. J., Eames, K. T. (2005). Networks and epidemic models. Journal of the Royal Society Interface, 2(4), 295-307.

Laguerre, Or, H. M. Hoang, and D. Flick. "Experimental investigation and modelling in the food cold chain: Thermal and quality evolution." *Trends in Food Science & Technology* 29, no. 2 (2013): 87-97.

Loaiza, J., and Cantwell, M. (1997). Postharvest physiology and quality of cilantro (Coriandrum sativum L.). HortScience, 32(1), 104-107.

Mabberley, D. J. (1997). The Plant-book: A Portable Dictionary of the Higher Plants. Cambridge: Cambridge University Press.

Manitz, J., Kneib, T., Schlather, M., Helbing, D., and Brockmann, D. (2014). Origin Detection during food-borne Disease Outbreaks-A case study of the 2011 EHEC/HUS Outbreak in Germany. PLoS currents, 6.

Marler Clark LLP. (2015). Foodborne Illness Online Database, 1984 – present.

McEntire, Jennifer, and Tejas Bhatt (2013). Pilot Projects for Improving Product Tracing along the Food Supply System–Final Report. Chicago, IL: Institute of Food Technologists (2013).

McGrayne, S. B. (2011). The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, & emerged triumphant from two centuries of controversy. Yale University Press.

McKellar, R.C., LeBlanc, D.I., Rodríguez, F.P., and Delaquis, P. (2014). Comparative simulation of Escherichia coli O157: H7 behaviour in packaged fresh-cut lettuce distributed in a typical Canadian supply chain in the summer and winter." Food Control 35, no. 1 (2014): 192-199.

Moore, C., and Newman, M. E. (2000). Epidemics and percolation in small-world networks. Physical Review E, 61(5), 5678.

Nesheim, M. C., Oria, M., & Yih, P. T. (Eds.). (2015). A framework for assessing effects of the food system. National Academies Press.

Newman, M. E. (2002). Spread of epidemic disease on networks. Physical review E, 66(1), 016128.

Newsome, R., Tran, N., Paoli, G.M., Jaykus, L.A., Tompkin, B., Miliotis, M., ... and Schaffner, D.W. (2009). Development of a Risk-Ranking Framework to Evaluate Potential High-Threat Microorganisms, Toxins, and Chemicals in Food. Journal of food science, 74(2), R39-R45.

Nsoesie, Elaine O., Sheryl A. Kluberg, and John S. Brownstein (2014). "Online reports of foodborne illness capture foods implicated in official foodborne outbreak reports." *Preventive medicine* 67: 264-269.

Nunn, L. H. (1981). An introduction to the literature of search theory. Professional Paper No. CNA-PP-305, Center For Naval Analyses (CNA) Corporation.
Nuzzo, J. B. (2013). When Good Food Goes Bad: Strengthening the US Response to Foodborne Disease Outbreaks. University of Pittsburgh Medical Center. Center for Biosecurity.

Nyachuba, David G. (2010). Foodborne illness: is it on the rise?. Nutrition reviews 68.5: 257-269.

Olson, Dean (2012). Agroterrorism: Threats to America's Economy and Food Supply. FBI Law

Onnela, J. P., Arbesman, S., González, M. C., Barabási, A. L., & Christakis, N. A. (2011). Geographic constraints on social network groups. *PLoS one*, *6*(4), e16939.

Osterholm, M.T. (2011). Foodborne Disease in 2011 — The Rest of the Story. N Engl J Med 2011; 364:889-891, March 10, 2011.

Painter, J.A., Ayers, T., Woodruff, R., Blanton, E., Perez, N., Hoekstra, R.M., ... and Braden, C. (2009). Recipes for foodborne outbreaks: a scheme for categorizing and grouping implicated foods. Foodborne pathogens and disease, 6(10), 1259-1264.

Painter, J.A., Hoekstra, R.M., Ayers, T., Tauxe, R.V., Braden, C.R., Angulo, F.J., Griffin, P.M. (2013). Attribution of foodborne illnesses, hospitalizations, and deaths to food commodities by using outbreak data, United States, 1998– 2008. Emerg Infect Dis 2013;19(3):407–15.

Pastor-Satorras, R., and Vespignani, A. (2001). Epidemic spreading in scale-free networks. Physical Review Letters, 86(14), 3200.

Pinior, B., Konschake, M., Platz, U., Thiele, H., Petersen, B., Conraths, F. C., Selhorst, T., (2012). The Trade network in the dairy industry and its implication for the spread of a contagion. Journal of Dairy Science, 95 (11), 6351-6361.

Pinto, P. C., Thiran, P., and Vetterli, M. (2012). Locating the source of diffusion in large-scale networks. Physical review letters, 109(6), 068702.

PMA (2013).Traceability and FSMA. http://www.pma.com/ resources/issues-monitoring/food-safety/fda-regulations-and-guidance/food-traceability-and-fsma.

Pouliot, S., Sumner, D. (2008). Traceability, Liability, and Incentives for Food Safety and Quality. American Journal of Agricultural Economics 90: 15–27.

Prakash, B. A., Vreeken, J., and Faloutsos, C. (2014). Efficiently spotting the starting points of an epidemic in a large graph. Knowledge and information systems, 38(1), 35-59.

Red Book Credit Services (2016). Accessed July, 2016. Available at http://www.rbcs.com

Reportbuyer. (2015). Global Agricultural and Environmental Diagnostics Industry. Accessed March, 2015.

Riley, S. (2007). Large-scale spatial-transmission models of infectious disease. Science, 316(5829), 1298-1301.

Scallan, E., Hoekstra, R. M., Angulo, F. J., Tauxe, R. V., Widdowson, M. A., Roy, S. L., ... and Griffin, P. M. (2011). Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis, 17(1).

Schoenfeld, A. (2007). A Multinational Loaf, The New York Times, June 15, 2007.

Seltzer, J. M., Rush, J., and Kinsey, J. D. (2009). Natural Selection: 2006 E. coli Recall of Fresh Spinach: A Case Study by The Food Industry Center (No. 54784). University of Minnesota, The Food Industry Center.

Shah, D., and Zaman, T. (2010). Detecting sources of computer viruses in networks: theory and experiment. In ACM SIGMETRICS Performance Evaluation Review (Vol. 38, No. 1, pp. 203-214). ACM.

Shah, D., and Zaman, T. (2012). Rumor centrality: a universal source detector. In ACM SIGMETRICS Performance Evaluation Review (Vol. 40, No. 1, pp. 199-210). ACM.

Smith, K., Miller, B., Vierk, K., Williams, I., and Hedberg, C. (2015). Product Tracing in Epidemiologic Investigations of Outbreaks due to Commercially Distributed Food Items – Utility, Application, and Considerations. Council to Improve Foodborne Outbreak Response (CIFOR). Accessed July, 2016. Available at:

http://www.cifor.us/clearinghouse/uploads/Product%20Tracing%20in%20Epidemiologic%20Investigations.pdf?CFID=10451937&CFTOKEN=36078807&jsessionid=33D02C90E2CA88ADF7D32DDA6DC4FC38.cfusion.

Smith, R., Bi, J., Cahn, M., Cantwell, M., Daugovish, O., Koike, S., Natwick, E., Takele, E. (2011). Cilantro production in California. UC Davis Agriculture and Natural Resources (ANR) Catalog, available at: http://anrcatalog.ucdavis.edu/pdf/7236.pdf

Stewart, S. R. (2013). "Tropical Cyclone Report: Tropical Storm Alvin." National Hurricane Center (National Oceanic and Atmospheric Administration). Published May 31, 2013. Retrieved December 1, 2014.

Storoy, J., Thakur, M., and Olsen, P. (2013). The Trace Food Framework - Principles and guidelines for implementing traceability in food value chains. Journal of food engineering 115(1):41-48.

This Week. FBI, 12 Sept. 2014. Web. 12 Mar. 2015.

USDA Agricultural Marketing Service (AMS). (2016). Market News Custom Reports. Accessed February 2016. Available at: https://www.marketnews.usda.gov/mnp/fv-report-config-step1?type=movement

USDA Economic Research Service (ERS) (2016). Retail Trends. Accessed February 2016. Available at http://www.ers.usda.gov/topics/food-markets-prices/retailing-wholesaling/retail-trends.aspx

USDA National Agriculture Statistics Service (NASS) (2012). Census of agriculture. Accessed February 2016. Available at https://www.agcensus.usda.gov/Publications/2012/Full_Report/Volume_1,_Chapter_2_US_State_Level/

USDA National Agriculture Statistics Service (NASS) (2016). Vegetables 2015 Summary. Accessed February, 2016. Available at http://usda.mannlib.cornell.edu/usda/current/VegeSumm/VegeSumm-02-04-2016.pdf.

Wang, X., Li, D., O'Brien, C., & Li, Y. (2010).A production planning model to reduce risk and improve operations management.International Journal of Production Economics124(2):463-474.

Wein, L.M. and Liu, Y.F. (2005) Analyzing a bioterror attack on the food supply: The case of botulinum toxin in milk. Proceedings of the National Academy of Sciences of the United States of America 102(28):9984-9989.

Wilkins, M., Julian, E., Kutzko, K., & Rockhill, S. (2015). Outbreak Investigations (Epidemiology). Regulatory Foundations for the Food Protection Professional, 105.

World Health Organization (WHO) (2011). Outbreaks of E. coli O104:H4 infection: update 30. July 7, 2011.http://www.euro.who.int/en/where-we-work/member-states/germany/sections/news/2011/07/outbreaks-of-e.-coli-o104h4-infection-update-30

World Health Organization. (2008). Foodborne disease outbreaks: guidelines for investigation and control. World Health Organization.

Wu, Y., Ranasinghe, D.C., Sheng, Q.Z., Zeadally, S., and Yu, J. (2011) RFID enabled traceability networks: a survey. Distributed and Parallel Databases 29(5):397-443.