A Universal Measure for Network Traceability

Xin Lu, Abigail L. Horn, Jiahao Su, Jiang Jiang

PII: S0305-0483(18)30129-4
DOI: https://doi.org/10.1016/j.omega.2018.09.004
Reference: OME 1960

To appear in: *Omega*

Received date: 5 February 2018
Revised date: 28 June 2018
Accepted date: 11 September 2018

Please cite this article as: Xin Lu, Abigail L. Horn, Jiahao Su, Jiang Jiang, A Universal Measure for Network Traceability, *Omega* (2018), doi: https://doi.org/10.1016/j.omega.2018.09.004

A Universal Measure for Network Traceability

Xin Lu,[1,2*] Abigail L. Horn,[3,4,5] Jiahao Su,[1] Jiang Jiang[1]

[1] *College of Systems Engineering, National University of Defense Technology, 410073 Changsha, China*
[2] *Department of Public Health Sciences, Karolinska Institutet, 17177 Stockholm, Sweden*
[3] *Engineering Systems Division, Massachusetts Institute of Technology, 02139 Cambridge, MA, USA*
[4] *Federal Institute for Risk Assessment (BfR), Max-Dohrn-Straße 8-10, 10589 Berlin, Germany*
[5] *Kühne Logistics University, 20457 Hamburg, Germany*

Abstract

In today's increasingly interconnected world, outbreaks of disease or contamination can spread widely and cause considerable impact on public health. Proactively assessing the ability to identify the source of an outbreak in a networked system is a critical step in aiding emergency and operational preparedness management prior to a crisis situation. While many methods have been developed to identify the source of an outbreak once it has occurred, limited research has been devoted to developing measures to assess the overall ability of a network structure to support accurate source identification, which we call *traceability*. Furthermore, while significant work has focused on understanding the role of network structure on propagation dynamics, its impact on traceability has yet remained unstudied. Here, we introduce a novel, comprehensive measure of network traceability, which calculates the information-theoretic entropy of the posterior probability distribution over feasible sources resulting from inferring the source location. By capturing information about the full posterior probability distribution, this measure presents an improvement over the binary logical outcome of the prediction accuracy metric generally applied to assess source identification method performance. Using food supply chain networks as an example, we use this measure to provide the first study systematically evaluating the role of network structural parameters on traceability, comparing both synthetic networks generated to exhibit a range of structural features known to be relevant to contamination propagation and real networks representing the Chinese pork supply chain across various cities. This analysis yields insights about the relationship between traceability and network structure, some counter-intuitive, and more generally, illustrates how this measure can be used in emergency and operational preparedness to proactively assess network traceability and recommend strategies for its improvement.

*Corresponding author. Email: xin.lu@flowminder.org

Keywords: traceability; entropy; complex networks; supply chain.

1 Introduction

An important problem for the management and regulation of networked systems involving spreading processes is improving the ability to trace the source of a spreading agent. This ability is of particular importance for both public health management and emergency preparedness in diverse contexts; if at the onset of a large-scale outbreak the contamination origin or source is identified efficiently, damage can be prevented or reduced [1, 2, 3, 4]. However, the ability to identify the source of an outbreak, which in this paper we refer to as *traceability*, can vary extensively for different networked structures and spreading processes. Proactively assessing network structural traceability can aid emergency and operational preparedness management on the need and allocation of resources to improve traceability.

In recent years much work has emerged on the problem of identifying the source of outbreaks spreading in a network. Approaches have been developed in the context of epidemic-type contagion processes including infectious disease outbreaks in human contact networks [5, 6, 7] or rumors spreading in social networks [8, 9, 10], and in transport-mediated diffusion-type processes such as diseases spread through water networks [11] and global air travel [12], or foodborne disease contamination spread through food distribution networks [13]. While the problem context and methodological approach vary, the general objective of the network source location problem is to rank all possible source nodes according to their relative likelihood of being the outbreak source, given available knowledge of the underlying network structure and the location (and in some cases the timing) of the reports of contamination at observing nodes. An intuitive framing of this problem, and that taken by the majority of the existing methods, is to invoke a probabilistic inference approach that determines a posterior probability that each possible source is the true source, and with it a maximum likelihood estimate of the true source location [8, 11, 6, 7, 13]. For a review of network source identification approaches classified by methodological approach, see [13].

Despite the substantial stream of work on the source location problem, limited research efforts have focused on developing quantitative measures of the overall ability of a network structure to support accurate source identification, i.e. *traceability*. When characterizing the accuracy of a proposed algorithm to identify the source of an outbreak of contamination or contagion on a specific network structure, the majority of the methods reviewed above report one of two metrics: mean prediction accuracy, based on the binary outcome of identifying the source in e.g. the first position or top-ten positions; or mean prediction rank, the position of the true source in the ordered ranking over all possible sources; both averaged over a large number of simulated outbreaks. These accuracy metrics

can be seen as the closest proxy to a quantitative measure of traceability in the existing literature. However when it comes to measuring the ability of a network structure to support accurate source identification, these accuracy metrics are suboptimal: relying on either a binary outcome or a single value (the rank of the true source), they do not take into account the rest of the information from the resulting posterior probability distribution over possible sources.

Furthermore, while the traceability of various network topologies has been investigated in order to demonstrate the robustness of an algorithm to differences in network structure, no research has contributed a systematic study of the role of network structural parameters in determining traceability. This is in spite of the well-established fact that network structure is a key determinant in spreading dynamics [14, 15, 16], with much research having focused on understanding how the properties of the underlying network shape the size and dynamics of potential spreading events. The seminal works in this area considered how spreading dynamics are shaped by the structural properties of stylized networks such as *small-world*, characterized by the property of high clustering and short average path lengths between any two nodes [17], and *scale-free*, characterized by the heterogeneous distribution of connectivity to other nodes [18, 19]. The influence of network structural properties such as degree heterogeneity, betweenness, degree density, weight distribution, and others on outbreak spreading characteristics have since been studied extensively over the past decade [20, 21, 22, 23]. More recently, this approach has been applied in the context of commodity or livestock trade with the aim of assessing how network topology mediates the dynamics and size of an outbreak of deliberate or unintentional contamination. Such work has investigated the role of heterogeneity in degree distribution, degree density, directionality, and modularity in swine and cattle trade networks [24, 25, 26] and the milk supply chain [27], and degree of overlap or mixing in the food production supply chain [28, 29].

Network structure should likewise be a determining factor in backward tracing. In the case of food distribution, one can easily imagine that for a supply network composed of vertically integrated supply chains, any observation of contamination can be correctly traced back to the original source [30]. 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. As the number of links increase, the number of pathways along which the contamination could travel will increase, and the predictability decreases. Each network structural variable may influence traceability in a particular way, which may change when taken together with other variables, allowing for factorial combinations.

1.1 Approach and Contributions

Studying the role of network structural parameters on traceability has important practical implications for emergency and operational preparedness, helping to develop an understanding of how and in what situations it might be possible to accurately identify an outbreak source. Being able to forecast where will be more difficult can inform the proactive allocation of resources to improve

network structural traceability. A systematic study of how differences in network structure impact this ability may therefore be of value to governmental agencies and regulatory bodies (e.g., the CDC, DOD, FDA, USDA) who may have influence on network design through regulations; or to industry groups who can take private market initiatives to adapt or influence network structure. 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.

This paper aims to address two gaps in the network source identification literature by (i) defining a novel, comprehensive measure for the traceability of a network structure, and (ii) using this measure to produce the first study that systematically evaluates the effect of network structural parameters on traceability. While the traceability measure itself is general and can be extended to any network source identification setting where a probability distribution is generated, we illustrate it here using the case of foodborne disease and food distribution networks. First we develop a stylized model of the food supply chain, a three-layered Farm-Distributor-Retailer model, and a food flow and mixing model. To estimate the source of an outbreak on this network, we adopt a Bayesian probabilistic inference approach, using the food flow and mixing model to identify the posterior probability distribution of any feasible source node being the true source in a manner similar to Horn and Friedrich [13]. Our contribution is then to define a measure of network traceability, the ability of a network structure to facilitate identification of the source, in a way that captures all information from the posterior probability distribution over all possible sources. This novel approach, network traceability entropy (NTE), calculates the information-theoretic entropy of the posterior probability distribution over possible sources, averaged over combinations of observed contaminated nodes. NTE thereby captures the full information provided by the differing probabilities while still generating a single output score for the *ability* of a network structure to support accurate source identification.

Entropy in information theory is a measure of the amount of information – or equivalently, the amount of uncertainty or unpredictability – carried by a random variable [31], in this case, the posterior probability distribution representing an estimate of the source location. While originally developed in communication theory to measure the how much useful information a message is expected to contain, the information-theoretic definition of entropy has, almost since its inception, been applied as a measure of system uncertainty, information, or predictability in many other fields such as in physics, mathematics, and statistics, as well as in complex system analysis in various domains from biology [32, 33, 34] to linguistics [35] to studies of human mobility [36, 37, 38]. It is thus particularly well suited to the task of measuring network traceability, since in this case it is calculated over the posterior probability distribution over possible sources and is measuring the predictability in identifying the true source.

By summarizing information about the full probability distribution over feasible sources into a single score, this measure presents an improvement over the existing simulation accuracy or rank

metrics that are based on a binary outcome or single rank value, while being just as convenient. In this paper we demonstrate the improvement, first comparing to the simpler simulation predictive accuracy metric and showing results are not 1-1. We also provide a simple comparative example to illustrate that there is more information captured in the entropy-defined measure than the binary predictive accuracy measure. A more comprehensive metric for network traceability can better inform policies to improve network traceability.

We then go on to use this measure to provide the first systematic comparison of network structures on their traceability. We identify a few network structural features that are important to spread and should likewise be important to traceability – link density, connective heterogeneity, and community structure. We parameterize these features and simulate network structures across a range of parameter values, then use NTE to draw conclusions about the role of network configurations on traceability. This comparative analysis using simulated network structures allows us to draw theoretical conclusions about the role of these specific network features on traceability. We then demonstrate the practical relevance of traceability analyses by including a case study based on data from China's new product tracing program, the "National Important Product Traceability System". We use this dataset to compare pork supply chain networks across 10 Chinese cities, and discuss how NTE might be used to recommend strategies for improving traceability in this industry.

Finally, we emphasize that in this paper we define the term *traceability* as a measure of the ability to identify the source of a spreading event in a network. This is a precise, quantitative definition of a network-theoretic measure. We acknowledge that the term has many uses and has received many definitions in the food safety, supply chain, technology development and regulatory literature. For a review of existing definitions of traceability and usage in technology development, see the Supporting Information.

1.2 Outline

The remainder of the paper is organized as follows. In Section 2, we define the food supply network and contaminations spreading models and define NTE, the measure of network traceability. In Section 3, we introduce the simulation setup and define Prediction Accuracy, a simulation-driven score of the accuracy of a identifying outbreak sources based on the binary accuracy metric used in the existing literature. In Section 4, we provide an illustrative example demonstrating how NTE efficiently encodes and transmits information regarding the uncertainty of the source identification problem. We compare NTE to the Prediction Accuracy measure as noted above. We then demonstrate how traceability entropy can be used to compare between various network configurations categorized by their link density, connective heterogeneity, and community structure. In Section 5 we apply the measure to network data on the pork supply chain in China. We discuss how NTE might be used to recommend strategies for improving traceability in this industry. Section 6 presents the conclusions drawn from our results and identifies extensions and future work.

2 Defining Network Traceability Using Entropy

2.1 Food supply network model

The food supply chain is composed of a wide diversity of products and companies which operate in different markets and sell a variety of food products. Here we will use a modeling framework that represents the network of distribution for a single commodity, such as spinach, but that can be generalized to any commodity. The steps encountered between production and consumption can vary considerably between food commodities as product moves through growers, processors, packagers, brokers, distributors, wholesalers, retailers, restaurants, etc. In order to use a model that is representative of the food supply chain in general, we aggregate the underlying trade network into the categories of farms, distributors, and retailers (Figure 1). In this simplified supply network, G, food is produced by each farm $F_i, 1 \leq i \leq |F|$, transported to different distributors $D_j, 1 \leq j \leq |D|$, mixed with food from other farms at these distributors, and then sold to customers at retailers $R_l, 1 \leq l \leq |R|$. Extensions to the three-layered network are straightforward. The number of farms, distributors, and retailers are denoted as $|F|$, $|D|$, $|R|$, respectively.

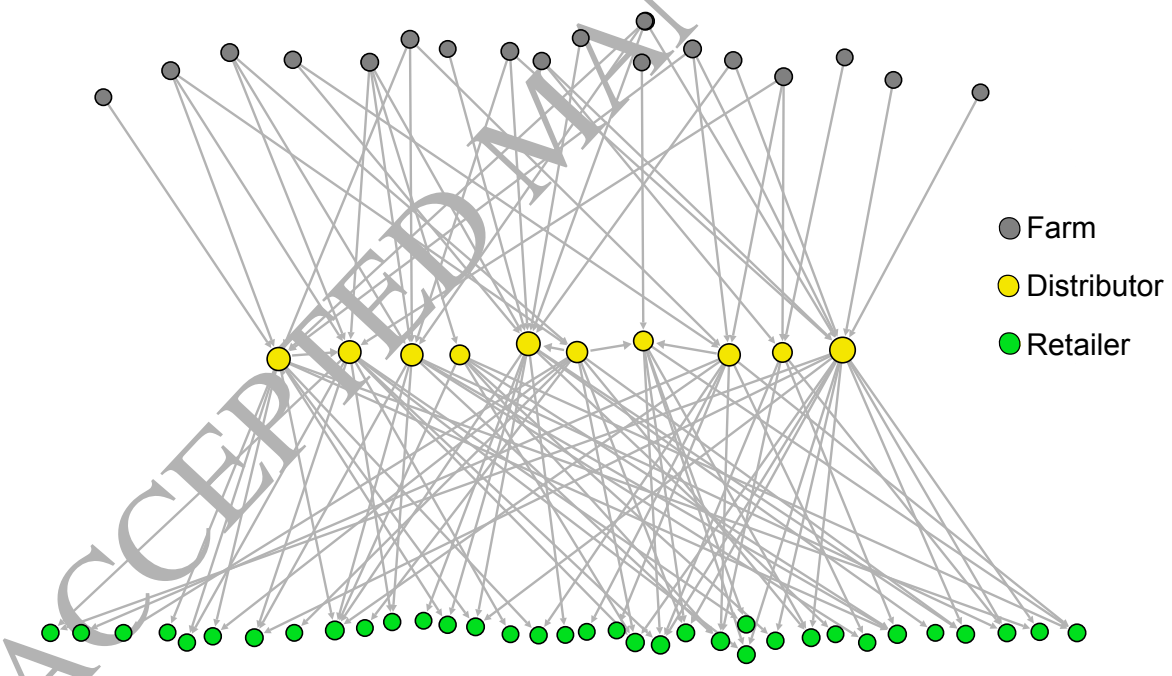

Figure 1: The Farm-Distributor-Retailer food supply network model.

2.2 Food flows and mixing

We now define a model for food flows through the supply network.

Let $f_i = x_i / \sum_{i'=1}^{|F|} x_{i'}$ be the proportion of food produced by farm F_i, $1 \leq i \leq |F|$, where x_i is

the total amount of food produced at farm F_i. Let FD, DD, DR be the food distribution matrices between the layers of farms, distributors and retailers, respectively. Let the elements in each of these matrix, fd_{ij}, dd_{jk}, dr_{kl} denotes the proportion of food products sent from the source node to the receiver node. Finally, let FR be the food composition matrix in which each element fr_{il} is equal to the proportion of food from farm F_i ending up at retailer R_l. FR can be calculated as:

$$FR = FD \times DD \times DR. \tag{1}$$

2.3 Probabilistic source identification

Given an outbreak of contamination reported at a set of retailer nodes $R_l^* \in \Omega, |\Omega| = \lambda$, the posterior probability that F_k is the true source can be written as

$$P(F_k|\Omega) = \frac{P(F_k)P(\Omega|F_k)}{\sum_{i=1}^{|F|} P(F_i)P(\Omega|F_i)}, \tag{2}$$

where $P(F_k)$ is a prior probability distribution over the k feasible source farms. If we assume that any unit of food produced is equally likely to generate contamination *a priori*. Then, the prior probability that any farm node is the source is equal to the relative production quantity at each source node, such that $P(F_k) = f_k, 1 \leq k \leq |F|$.

To determine the likelihood, we first expand $P(\Omega|F_k)$ in terms of the individual illness reports at nodes $R_l^* \in \Omega$,

$$P(\Omega|F_k) = P(R_1^*, R_2^*, \ldots, R_\lambda^*|F_k). \tag{3}$$

If we assume that each illness observation i is mutually independent of every other observation, the joint probability distribution in Equation 3 can be factorized into a product of the individual probabilities of each reporting retailer R_l^*, such that

$$\begin{aligned} &P(R_1^*, R_2^*, \ldots, R_\lambda^*|F_k) \\ &= P(R_1^*|F_k)P(R_2^*|F_k)\ldots P(R_\lambda^*|F_k) = \prod_{R_l^* \in \Omega} P(R_l^*|F_k). \end{aligned} \tag{4}$$

In practice, we can reasonably expect the condition of mutual independence to be validated [39]. This can be understood by noting that the observations of illness at nodes R_i are mutually independent if all nodes receive contaminated food from batches departing separately from the source. For large contamination incidents where the contaminated quantity will be larger than what fits in one truck, this is necessarily the case. Furthermore, large-scale outbreaks are characterized by the widespread distribution of contaminated product throughout the supply chain resulting in reports of contamination in disperse geographical locations [40, 41]. This dispersion implies that many more contaminated food items leading to reports of illness at different nodes will travel in separate batches than will travel together.

If we assume that the probability of reporting infection is proportional to the amount of contaminated food received at each retailer, then given an outbreak at farm F_k, the probability that retailer node R_l^* observes an illness is then equal to the proportion of food produced at farm F_k delivered to retailer R_l^*, i.e.

$$P(R_l^*|F_k) = fr_{kl}, \tag{5}$$

The likelihood is consequently found as the product of the proportions delivered to each $R_l^* \in \Omega$,

$$P(\Omega|F_k) = \prod_{R_l^* \in \Omega} fr_{kl}, \tag{6}$$

And equation 2 becomes

$$P(F_k|\Omega) = \frac{f_k \prod\limits_{R_l^* \in \Omega} fr_{kl}}{\sum\limits_{i=1}^{|F|} f_i \prod\limits_{R_l^* \in \Omega} fr_{il}}. \tag{7}$$

Then for each set of observations Ω, we obtain a list of posterior probabilities: $P(F_1|\Omega), P(F_2|\Omega), \ldots, P(F_k|\Omega), \ldots, P(F_{|F|}|\Omega)$ which forms a distribution and fulfills $\sum\limits_{k=1}^{|F|} P(F_k|\Omega) = 1$.

2.4 Network Traceability Entropy

With the posterior probability distribution defined, we can use the information theoretic definition of entropy as a measure of the uncertainty in determining the source farm given an outbreak in the node set Ω,

$$E(\Omega) = -\sum_{k=1}^{|F|} P(F_k|\Omega) \log P(F_k|\Omega). \tag{8}$$

Entropy can be seen as a measure of uncertainty about a value sampled from a probability distribution [31, 36, 37]; in this case, it measures the uncertainty in determining the source of the outbreak F_k given the contamination set Ω. If we consider a game where we are allowed to ask yes/no questions until we identify the correct outbreak source, the entropy of the posterior probability distribution $P(F_k|\Omega)$ can be interpreted as the average number of (suitable) yes/no questions we need to ask to pinpoint the source. Turning this around, 2^E is the uncertainty in the source location, or the *effective number* of plausible source candidates. The reader may wish to refer to Section 4.1 for an example illustrating this principle.

If we have little uncertainty about the source, then we get to the correct answer with a few questions. An extreme case is when $E(\Omega) = 0$. This can happen when only one feasible source node F_k exists, with probability $P(F_k|\Omega)$. In this case we do not need to ask any questions to get to the

value of the true source. In the other extreme, uncertainty is greatest when the amount of food is distributed equally across all farms to all retailers, i.e. $E(\Omega) = -\sum_{k=1}^{|F|} \frac{1}{|F|} \log \frac{1}{|F|} = \log |F|$. These two extreme cases bound $E(\Omega)$ as

$$0 \leq E(\Omega) \leq \log |F|. \tag{9}$$

Finally, we define the *Network Traceability Entropy (NTE)* as the average entropy for food supply network G given any number of contamination reports λ:

$$E^{\lambda} = \sum_{\Omega:|\Omega|=\lambda} E(\Omega) / \binom{|R|}{\lambda}. \tag{10}$$

Network Traceability Entropy so defined permits an intuitive definition of the uncertainty of source traceback. While food distribution networks are used as the illustrative example in this paper, the measure can apply to any problem of network source identification when a posterior probability distribution for the source location can be computed given some number λ of reported cases.

3 Experimental Design

We have proposed NTE as a measure of the uncertainty of accurate source identification in a network, *traceability*. We now introduce a simulation environment that we will use to evaluate the measure and present a systematic study of the role of network structure on traceability.

3.1 Network configuration

We adopt a few key structural features to generate a spectrum of networks across varying parameters (Table 2). First, to approximate the relative size of food supply network actors, we fix the number of farms, distributors, and retailers at $|F| = 100$, $|D| = 20$ and $|R| = 500$, respectively. Let $\bar{d}_F^{out}$, $\bar{d}_{FD}^{in}$, $\bar{d}_{DR}^{out}$, and $\bar{d}_R^{in}$ denote the average outdegree of farms, average indegree from farms to distributors, average outdegree from distributors to retailers, and average indegree of retailers from distributors, respectively. Based on $\bar{d}_F^{out}$, $\bar{d}_{FD}^{in}$, $\bar{d}_{DR}^{out}$, and $\bar{d}_R^{in}$, we generate sets of degrees d drawn from certain probability distributions. We then randomly pair links between related layers according to the Network Configuration Model [42] . In the following studies we fix the degree distributions to be uniform ($d \sim U(1, 2\bar{d}-1)$), binomial ($d \sim B(d_{max}, \bar{d}/d_{max}$))), or exponential ($d \sim Exp(\bar{d})$), and vary the density of connections $\bar{d}_F^{out}$ and $\bar{d}_{DR}^{out}$ from *low*: $\bar{d}_F^{out} = 2$, $\bar{d}_{DR}^{out} = 30$; *medium*: $\bar{d}_F^{out} = 3$, $\bar{d}_{DR}^{out} = 40$; to *high*: $\bar{d}_F^{out} = 4$, $\bar{d}_{DR}^{out} = 80$. We choose these values because they are representative of real food supply networks, as can be seen by comparison to the 10 real networks from the case study analysis in Section 5 (see Table 2).

Unless otherwise stated, we assume all farms produce a unit amount of food and for each link food is distributed evenly among all outgoing links of an actor. For the purpose of this study, we consider only links between farms and distributors, and distributors and retailers. It is straightforward to extend the approach demonstrated to networks with more complicated interactions.

3.2 Simulation setup

With the above specifications, we generate networks for each of the *low*, *medium* and *high* settings, with degree distributions configured from uniform, binomial and exponential distributions. For each density and degree distribution setting we generate 1000 networks randomly, for a total of 9000 networks. The *NTE*, E^{λ}, is calculated for each network according to Equation 10.

3.3 Prediction accuracy benchmark

The closest proxy to a quantitative measure of traceability in the existing literature is the standard predictive accuracy metric used to assess the mean source identification algorithmic performance for a given network structure: an outbreak is simulated, the source identification algorithm is applied, and a binary logical outcome recorded representing whether the simulated source is identified. When this accuracy metric is averaged over a large number of simulated outbreaks for a given network structure, this can be considered a measure of network traceability. Here we formally define such a measure, *prediction accuracy*, $\langle\gamma(g)\rangle$, in order to provide a benchmark of the state-of-the-art measure of network traceability, which we will use to make comparisons with *NTE*.

Spreading scenarios are generated through Monte Carlo simulation. First, a farm F_s is randomly chosen as the source of contamination and the subset, Θ, of all retailers that can be reached by this farm is determined. We assume that only a subset of retailers in Θ actually present with the contamination, a result of the contamination being passed between links with a probability less than one, as well as example-specific factors such as underreporting of illness. This subset, Ω, is sampled in proportion to the food composition fr_{si} received at each retailer R_i reached by the source farm. The probability that a farm F_k is the outbreak source given the reports of illness at retailers $R_i \in \Omega$ can then be determined by applying the source estimator in Equation 7. To compute the prediction accuracy for a given network, we simulate 1000 iterations of the infection and prediction process. For each iteration we allow a variable number of guesses, g, for the source farm, assigning a 1 if the outbreak source F_s is within the top g farms with highest probability values and a 0 if it is not. $\langle\gamma(g)\rangle$ is calculated as the fraction of correct predictions.

4 Results

4.1 An illustration of the calculation of *NTE*

In this section we apply *NTE* to a specific outbreak example to help establish the reader's practical understanding of the measure. The example furthermore illustrates, explicitly, how traceability entropy effectively encodes information about the uncertainty of the source identification problem.

Figure 2 plots the probability distribution $P(F_k|\Omega)$ for the outbreak source farm resulting from a contamination event given $|\Omega| = 1$, 2, and 3 reports of illness. The probability values are ordered by their arbitrary numerical identifier, i.e. *farm ID*. The example is simulated on a network in which all degree distributions are binomial, $d \sim B(d_{max}, \frac{\bar{d}}{d_{max}})$, and the density of connections are *medium*, i.e. $\bar{d}_F^{out} = 3$ and $\bar{d}_{DR}^{out} = 40$. For reference, the outbreak in this simulation was generated from *farm ID 21*.

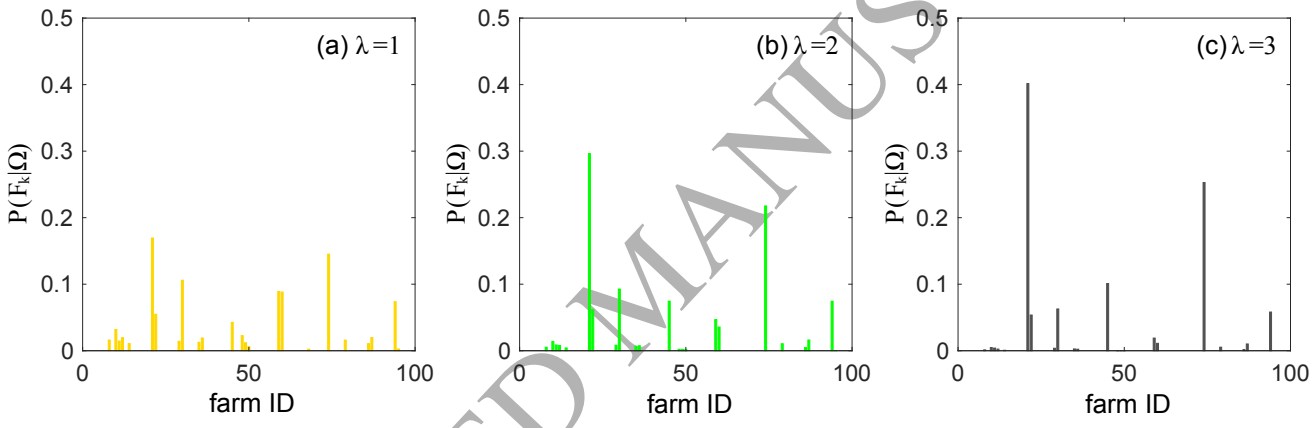

Figure 2: Example distribution for the probability of being an outbreak source given (a) $|\Omega| = 1$; (b) $|\Omega| = 2$; and (c) $|\Omega| = 3$.

The example in Figure 2 demonstrates expected behavior: as the number of observations of illness increase, the uncertainty in identifying the source farm will decrease. This is due to two important factors: fewer sources are *possible* options, and, fewer sources *probable*. First, additional observations will rule out farms that are unable to reach all retailers in the observation set. In the example we see that the number of feasible farms, or farms with a nonzero probability, decreases from 26 to 22 as the reports increase from $|\Omega| = 1$ to $|\Omega| = 3$. Second, the probability values distributed across the farms will become more pronounced as the number of terms in the product forming the likelihood function increase. This differentiation of probability values is clearly visible as the number of reports increase over the 3 plots in Figure 2, with the low values becoming lower while the high probability values grow higher. This is especially noticeable for the maximum probability values, the true source at *farm ID 21*.

Table 1: Illustrative calculations for the examples presented in Figure 2.

#. observations $\lvert\Omega\rvert$	Entropy $E(\Omega)$	#.feasible source	*effective* #. feasible sources 2^E	#. farms with $P(F_k \mid \lvert\Omega\rvert = 2) > 0.05$
1	3.9	24	14.6	13
2	3.2	23	9.1	8
3	2.6	22	6.0	6

While the probability distribution across all feasible sources is necessary to fully characterize the uncertainty in the source identification problem, we now show how the traceability entropy efficiently encodes and transmits information about this uncertainty. Although we do not have an intuitive understanding of the meaning of the entropy score itself, the reverse value of 2^E, as discussed in previously, has a practical meaning we can understand: it represents the uncertainty in the source location, or the *effective number* of feasible sources. This can be reflected in what we see in the distributions plotted in Figure 2. Table 1 summarizes the entropy score, the number of feasible source candidates, and the *effective number* of feasible source candidates for the 3 scenarios in the example. When $|\Omega| = 2$, the number of feasible farms is 24 while the *effective number* is $2^E = 9.1$, which corresponds roughly to the 8 farms in Figure 2(b) with significant probability values, e.g., $P(F_k \mid |\Omega| = 2) > 0.05$. When $|\Omega| = 3$, the number of feasible farms has decreased only by 2, but the *effective number* of feasible farms has decreased to $2^E = 6$, which corresponds exactly to the 6 farms with significant probability values. While not every calculation of 2^E will correspond so precisely to the *effective number* of feasible farms, this example functions to demonstrate the useful relationship between entropy and uncertainty in the source detection problem.

To take this example one step further, in Figure 3 we show simulation results for E^1, E^2 and E^3 for the same network setup. The E^λ results are averaged across 100 simulated network structures with the error bar designating one standard deviation from the mean. We observe that E^1 is around 4.1, indicating that on average for this network, the *effective number* of feasible contamination sources is $2^{4.1} = 17.1$ farms after one illness has been observed. Increasing the number of observations greatly reduces the uncertainty in this network: when two illnesses have been observed, the *effective number* of feasible sources reduces from 17.1 to $2^2 = 4$ farms; with three observations, it reduces to $2^{0.6} = 1.5$, fewer than two farms.

4.2 Noting the difference between *NTE* and *prediction accuracy*

We now provide a simple example to illustrate how *NTE* captures more information than the binary *prediction accuracy* metric defined in Section 3.3, which can be seen as the state-of-the-art measure of network traceability. Imagine a scenario involving two networks, A and B, each with three farms and one retailer. Given a contamination observed at the retailer in each network, let's suppose

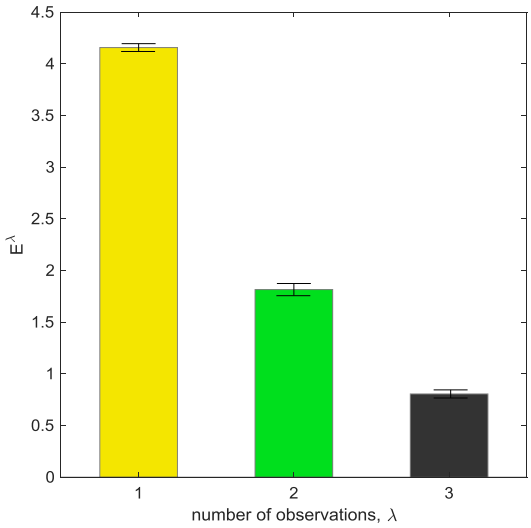

Figure 3: Illustration of traceability entropy E^1, E^2 and E^3 for a network in which all degree distributions are binomial, $d \sim B(d_{max}, \frac{\bar{d}}{d_{max}})$, and the density of connections are *medium*, i.e. $\bar{d}_F^{out} = 3$ and $\bar{d}_{DR}^{out} = 40$. Results are averaged across 100 simulated network structures with the error bar designating one standard deviation from the mean.

that the probability of identifying the source is found to be equal to the distribution $P_A(F \mid \Omega) = \{0.5, 0.49, 0.01\}$ for A and $P_B(F \mid \Omega) = \{0.5, 0.25, 0.25\}$ for B. The *NTE* for A and B in this scenario are $E_A^1 = 0.7422$ and $E_B^1 = 1.0397$, which indicates that the ability to identify the source of outbreaks in network A is more predictable, or that network A encodes less uncertainty, than network B. However, if we apply the prediction accuracy metric with $g = 1$, we can see that the prediction rate would be $\langle \gamma_A(g) \rangle = \langle \gamma_B(g) \rangle = 0.5$; that's to say, if we use prediction accuracy to measure the ability of tracing, the two networks becomes indistinguishable.

4.3 Correlation between *NTE* and prediction accuracy

In this section we quantify the correlation between the entropy-based measure and the prediction accuracy metric, demonstrating that the measures are not identical. As demonstrated in the examples above, *NTE* is calculated based on the full posterior probability distribution over all sources, encoding more information than the binary prediction accuracy metric.

To compare *NTE* and the prediction accuracy, we compute both measures in parallel for each of the 9000 networks generated for evaluation. Figure 4 presents the results, demonstrating a strong correlation between the *NTE* E^λ and the prediction Accuracy $\langle \gamma(g) \rangle$. As expected, the accuracy of identifying the source of simulated outbreaks decreases as *NTE* increases. With one guess, prediction accuracy is a monotonically decreasing function of *NTE* with strong correlation; the Pearson corre-

lation coefficient for both $\lambda = 1$ (Figure 4(a)) and $\lambda = 2$ (Figure 4(c)) is as high as -0.97 (significance $p < 0.000$). Prediction accuracy clearly increases with the number of guesses, i.e., the more farms that can be investigated, the more likely it is to find the correct contamination source. For these stylized networks, when $g = 10$ and $\lambda = 1$, there is a $60\% \sim 70\%$ chance of finding the source for networks with low *NTE* ($E^1 < 3$), while the likelihood remains low ($30\% \sim 40\%$) for networks with high *NTE* ($E^1 > 4$).

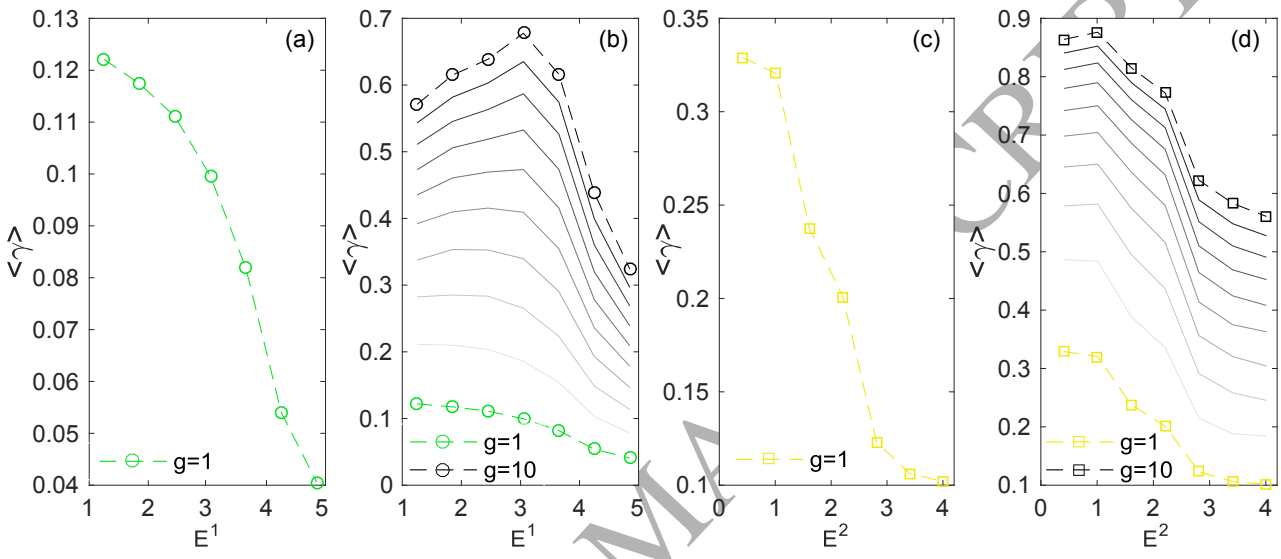

Figure 4: Prediction accuracy $\langle\gamma(g)\rangle$ for different numbers of guesses g as a function of *Network Traceability Entropy* E^λ, based on outbreaks observed in one retailer (a-b) and two retailers (c-d). In (a) and (c), the prediction accuracy $\langle\gamma(1)\rangle$ for $g = 1$ guess for the source farm is highlighted. In (b) and (d), intermediate curves represent the number of guesses g from 2 to 9.

Still, there are areas where the correlation between *NTE* and the *Prediction Accuracy* is not monotonic, which is attributable to the differences in the two measures as discussed above.

4.4 Effect of Network Structure on Traceability

In this section we use Network Traceability Entropy to facilitate comparisons between the traceability of different network configurations. Food distribution networks are characterized by multiple structural features, which may exert a distinct influence on traceability either independently or in combination. We apply the measure to evaluate the role of several structural parameters on traceability: degree density, degree heterogeneity, and regional structure.

4.4.1 Network density and degree distribution

The density and the heterogeneity of connections between farms, distributors and retailers are two major factors affecting traceability. High link density means more pathways for disseminating the contamination; the number of pathways along which the contamination could travel will increase, and the predictability decreases. Furthermore, the distribution of the degrees supplies relevant information about the structure of a network and also about food safety. Food distribution networks are characterized by heterogeneity in (i) the distribution of the number of links in or out of each node across all nodes in a stage, or the in- or out-degree distributions, respectively; (ii) the distribution of flow volumes across the links leaving a single node; and (iii) the initial volume distribution across producing nodes at the first stage. This behavior has been observed in network studies documenting supply chain structure [43, 44, 27, 45], and moreover is characteristic of complex networks in general [12, 46].

To investigate the role of degree density and heterogeneity on *NTE*, we generate networks for each of the *low*, *medium* and *high* degree density settings with degrees configured according to uniform, binomial and exponential distributions, for a total of nine separate network settings. The *NTE* scores E^1 and E^2 are then calculated as the average over 100 randomly generated networks for each density and degree setting.

The results demonstrate that *NTE* follows expected properties, reflecting changes in the traceability resulting from variations in network complexity (Figure 5). First, we observe that *NTE* visibly increases with network density for both one and two observations. This follows our expectations, since greater connectivity means more paths for the contamination to have traveled, more farms to be feasible sources, and correspondingly more uncertainty in backward tracing. For example, when all degrees are drawn from a binomial distribution and with two observations, the traceability entropy increases from 0.5 when the network is sparse (*low*), to 3 when the network becomes dense (*high*). These values correspond to an increase in the *effective number* of feasible sources of $2^{0.5} = 1.4$ for the sparse configuration and $2^3 = 8$ for the dense configuration, corresponding to a decrease in the ability to accurately pinpoint the source.

Second, with regard to the effect of the degree distribution heterogeneity, we observe that *NTE* scores for networks generated from a binomial distribution are higher than for those generated from uniform and exponential distributions. Values sampled from a binomial distribution will be focused around a mean value whereas values sampled from uniform and exponential distributions will be more heterogeneously distributed. As a result, the amount of food will be more evenly distributed between farms and retailers in the binomial networks. The evenly spread distribution of food corresponds to less probabilistic distinguishability between pathways the contamination could have traveled, and increased uncertainty in the backward tracing problem.

Third, we observe that for each degree distribution setting, the decrease in *NTE* from one observation (E^1) to two observations (E^2) becomes less distinguishable for higher link density settings.

This again follows our expectations, since the effect of an additional observation of illness in a highly connected food distribution network will correspond to a less significant reduction in the number of feasible pathways and sources from which the contamination could have traveled, and thus a less significant reduction in uncertainty in identifying the source.

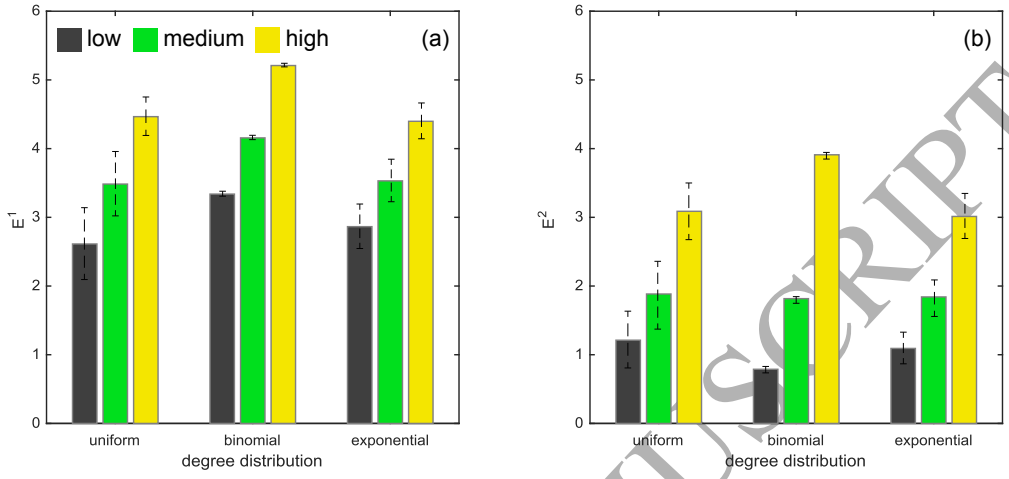

Figure 5: *Network Traceability Entropy* for food distribution networks with different densities and degree distributions. (a) E^1; (b) E^2. Grey, green, and yellow indicate *low*, *medium*, and *high* degree settings, respectively. Error bars represents one standard deviation from the mean for each entropy calculation.

4.4.2 Joint effect of degree distributions

We have now seen that for networks generated exclusively from uniform, binomial, or exponential distributions, the shape of the distribution affects the traceability of the network. In this section, we further explore the relationship between degree distribution and traceability by systematically evaluating *NTE* for multiple combinations of in- and out-degree patterns. Specifically, we will consider the $3^4 = 81$ combinations of uniform, binomial, and exponential degree distributions from the 4-dimensional parameter space $d_F^{out}, d_{FD}^{in}, d_{DR}^{out}, d_R^{in}$. As in the basic setup, the average degree is held constant for each distribution. The *NTE* for $\lambda = 1$ observation is calculated as the average over 100 randomly generated networks for each of the 81 settings.

Figure 6 presents the results in what we call a *fractal heat map*. The map is composed of nine major cells each of which is again composed of nine minor cells, allowing the four dimensional space to be represented in a two-dimensional planar plot. Major cells represent the out-degree of farms (d_F^{out}) and in-degree of distributors (d_{FD}^{in}) and minor cells represent the out-degree of distributors (d_{DR}^{out}) and in-degree of retailers (d_{DR}^{in}), with each of the three distributions (binomial, uniform, exponential) represented as indicated.

Consistent with findings in the previous section, we can see in Figure 6(a) that the highest values for traceability entropy are observed when the network is most homogenous, i.e., when all degree distributions are binomial. Between the out-degree distribution of farms and the in-degree distribution of distributors, the out-degree dominates the effect on the traceability entropy: low entropy occurs only when the out-degree distribution of farms is uniform or exponential, and reaches its lowest values when the out-degree distribution of farms is uniform and the out-degree distribution of distributors is uniform or exponential. These results emphasize the importance of the degree configurations on traceability, demonstrating the need to apply network traceability studies to evaluate multi-factorial combinations of feature dimensions.

We have also investigated the effect of degree distributions under conditions of inhomogeneous initial production quantities across the farms. Instead of dividing the proportion of food evenly across all farms, we divide the proportion of food evenly across all links. The *Network Traceability Entropy* results for the 81 network configurations are presented in the *fractal heat map* in Figure 6(b). As can be seen by comparing Figure 6(a) and (b), the relationship between the degree distributions and traceability are quite similar. When each link carries the same quantity of food, *NTE* scores marginally increase in magnitude, and the patterns of difference between distribution types become slightly amplified. The observed consistency implies that traceability depends more on the structure of the supply network than on the initial conditions set by the quantity of food produced at each farm. This is an important takeaway and should be explored further in future research on network traceability.

4.4.3 Regional effects

Two major trends have influenced the development of food market structure in recent years, and they are polarizing: centralization and regionalization. The global industrial food system has undergone dramatic centralization, or the concentration of production into fewer, larger actors with more connections and larger supply reach [47, 48] . This trend has obvious economic benefits, increasing efficiency, reducing costs, and producing higher profit margins. The trend is also motivated by the consumer demand for a wider availability of food products, which have more than doubled between 2000 and 2010 [27] . At the same time, locally produced food is seeing a reemerging demand, mainly in developed countries, for a variety of reasons including higher product quality and stimulation of local economy [48, 49] .

To model regional structure in food supply networks, we evenly divide the farm, distributor, and retailer supply chain sectors into five hypothetical regions. We introduce a parameter p determining the adherence to these regions, such that links are formed between supply sectors within the same region with probability p and with actors across all regions with probability $1 - p$. Consequently, $p = 0$ represents the absence of regional structure, or complete centralization, and $p = 1$ represents the extreme constraint that food is only distributed within the local community, or complete region-

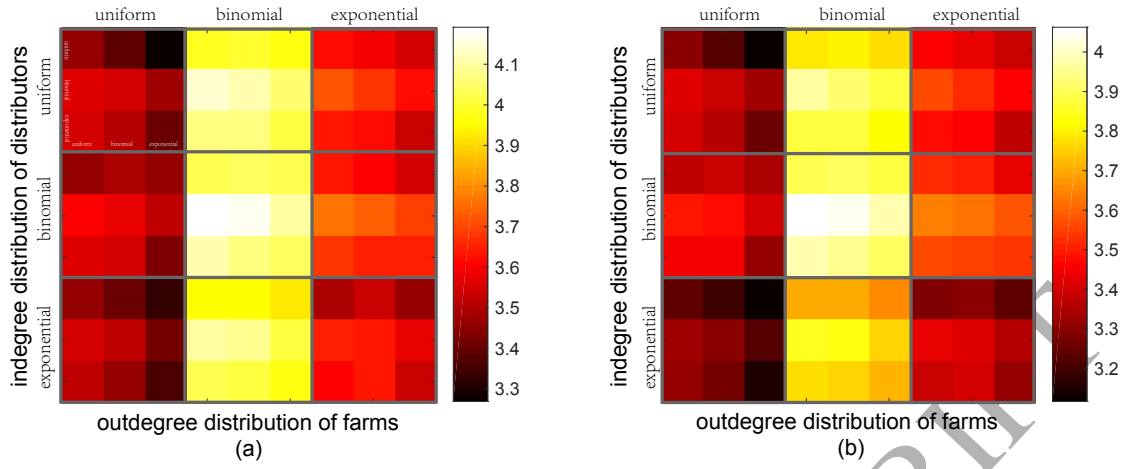

Figure 6: *Fractal heat map* depicting *Network Traceability Entropy* E^1 results for networks generated with different distributions determining the configurations of links out of and into each supply stage of farms, distributors, and retailers, i.e. $d_F^{out}, d_{FD}^{in}, d_{DR}^{out}, d_R^{in}$. (a) All farms produce same amount of food; (b) the amount of food produced in each farm is proportional to its out-degree. Major grid cells in the heat maps represent the out-degree of farms (d_F^{out}) and in-degree of distributors (d_{FD}^{in}) and minor cells represent the out-degree of distributors (d_{DR}^{out}) and in-degree of retailers (d_{DR}^{in}), with each of the three distributions (binomial, uniform, exponential) represented as indicated.

alization. We then test the effect of p on *NTE* at each regionalization value and each density, for a total of 18 network settings. Results are presented in Figure 7.

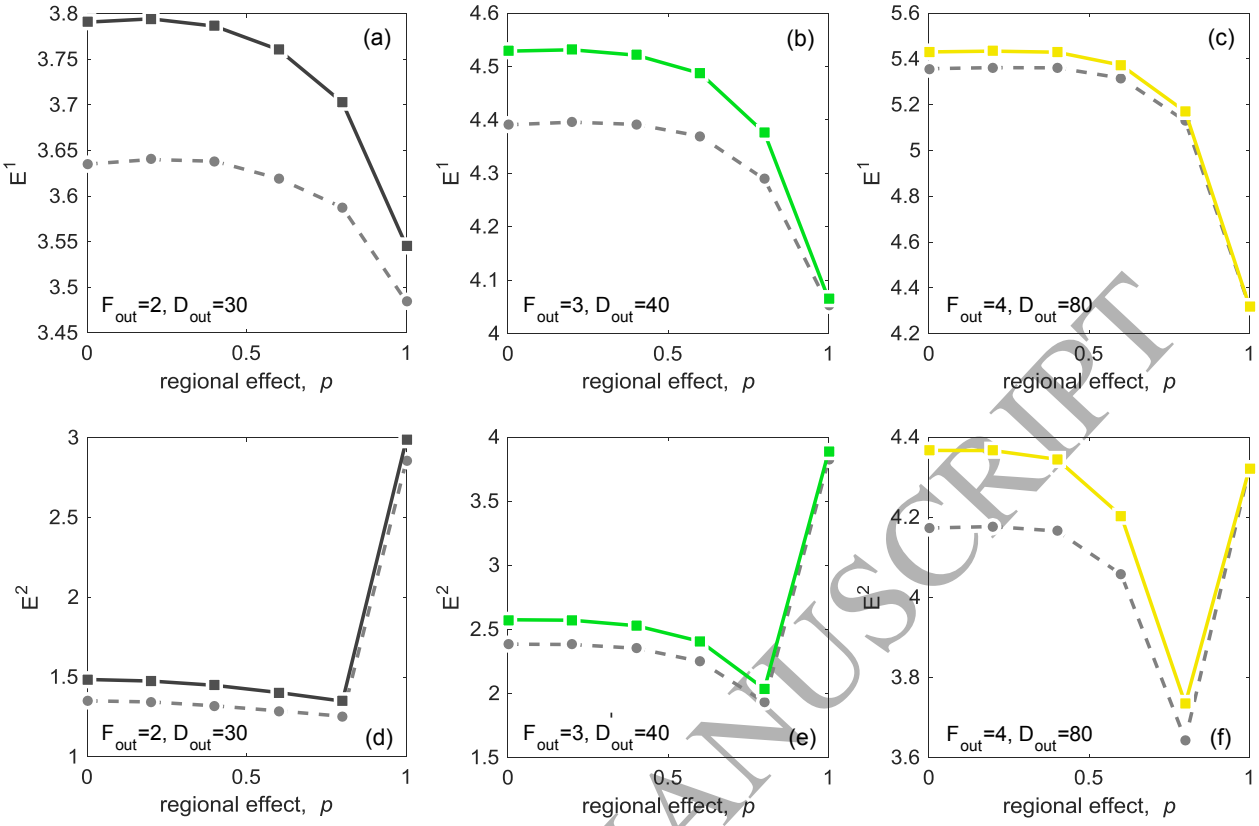

Figure 7: *Network Traceability Entropy* as a function of regional structure. (a) - (c) E^1; (d) - (f) E^2. From left to right grey, green, and yellow indicate *low*, *medium*, and *high* density, respectively. Each plot (a) through (f) depicts the traceability entropy E^λ for networks generated with the regional constraint parameter p, for the given density level. Solid lines represent food distributed equally across the farms and dashed lines represent food distributed equally across the outgoing degrees.

We observe that when only one node has reported illness, *NTE* decreases as local structure increases; in other words, as centralization decreases, traceability increases. This is as expected, since the number of sources any single retailer can connect to will decrease as the regional components become more segregated; in the extreme case, a given retailer will only be able to connect to the subset of farms within its region.

When two nodes have been contaminated, traceability entropy exhibits notably different behavior, decreasing with increasing p but jumping up to a maximum value when p reaches 1 and the regions have become segregated. What is happening is that when $p = 1$, independent sub-networks have been created that are smaller yet denser than the original network. When $p < 1$, *NTE* gradually decreases as the localization constraint increases for similar reasons as for one contaminated retailer; either both contaminated nodes are in the same region and can connect to the subset of farm nodes within that region, or the nodes are in different regions and have an even lower chance of linking to

farms connecting to both regions.

We also note that the above findings are not biased by the initial distribution of food production. The effect of regionalization on E^1 and E^2 is very similar whether the amount of food produced is evenly distributed across the farms or evenly distributed across all outgoing links.

These findings suggest that regional structure has a significant effect on a network's traceability. For non-extreme cases, i.e. regionality parameter values around 0.5, cross-regional structures with high centralization exhibit lower traceability than less dense, more contained local structures. However decreasing centralization and increasing local supply structure will improve traceability only to a point. When entire sections of a network are isolated into independent components, a threshold is crossed and traceability decreases. Some degree of mixing between components is needed to limit the feasible source set and decrease the uncertainty in the source identification problem. This insight has important implications for policy and practice and should be further investigated given current trends in global food market systems.

5 Case Study

5.1 Data description

Our case data comes from China's "National Important Product Traceability System" (NIPTS), which is initiated under the instructions of the State Council and is constructed by the Ministry of Commerce (MOFCOM), in conjunction with Ministry of Industry and Information Technology (MIIT), Ministry of Agriculture (MOA), China's Food and Drug Administration (CFDA), among other agencies, to build a national level platform for sharing and exchange of traceability data amongst relevant departments, enterprises, producers and traders of key products including edible agriculture products, food products, drugs, agricultural inputs, special equipment, and hazardous materials and rare earth. The NIPTS was established in 2014 with the goal of improving food safety and promoting the government's "Internet and Agriculture Act" [50]. The system has been implemented in more than 15,000 corporations and 320,000 shops across 58 piloting cities, and continues to expand. Since its establishment in 2014, more than 2 billion transaction records have been collected, with the NIPTS receiving around 3 million records on a daily basis. Details of the implementation and updated progress of the NIPTS can be found from the government's webpage www.zyczs.gov.cn.

Without loss of generality, we analyze pork supply chain data from the first batch of 10 NIPTS pilot cities (marked as green nodes in Figure 8) during the year of 2015. The data documents the flow of pork between the three supply chain stages of farms, slaughterhouses, and retailers for participating establishments. For each establishment, the origin, destination, date, and volume of each transaction are recorded. Setting farms, slaughterhouses, and retailers as nodes and transactions

as links, we transform the data into supply chain networks. As the NIPTS is still expanding, the data and resulting networks may only cover a portion of the real pork supply chain in the cities considered. The analysis demonstrated here therefore serves as an illustration of what can be learned by applying *Network Traceability Entropy* to real network data, and how it can inform the design or redesign of networks for improved traceability.

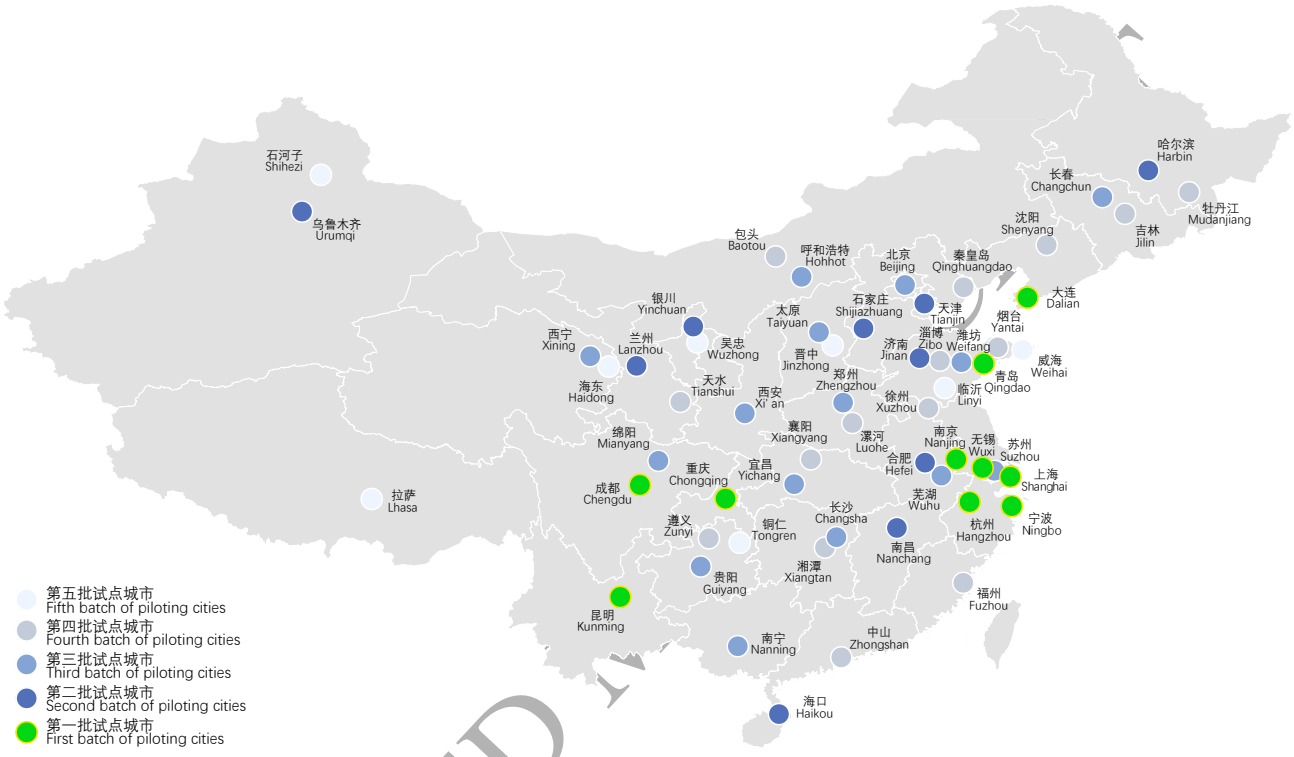

Figure 8: Map of cities covered by the "Important Product Traceability System" (NIPTS), a Chinese Ministry of Commerce initiative established in 2014. As part of this system, traceability data on a variety of food products has been collected across 50 piloting cities as indicated. We study pork supply chain data for the first batch of 10 piloting cities, marked in green.

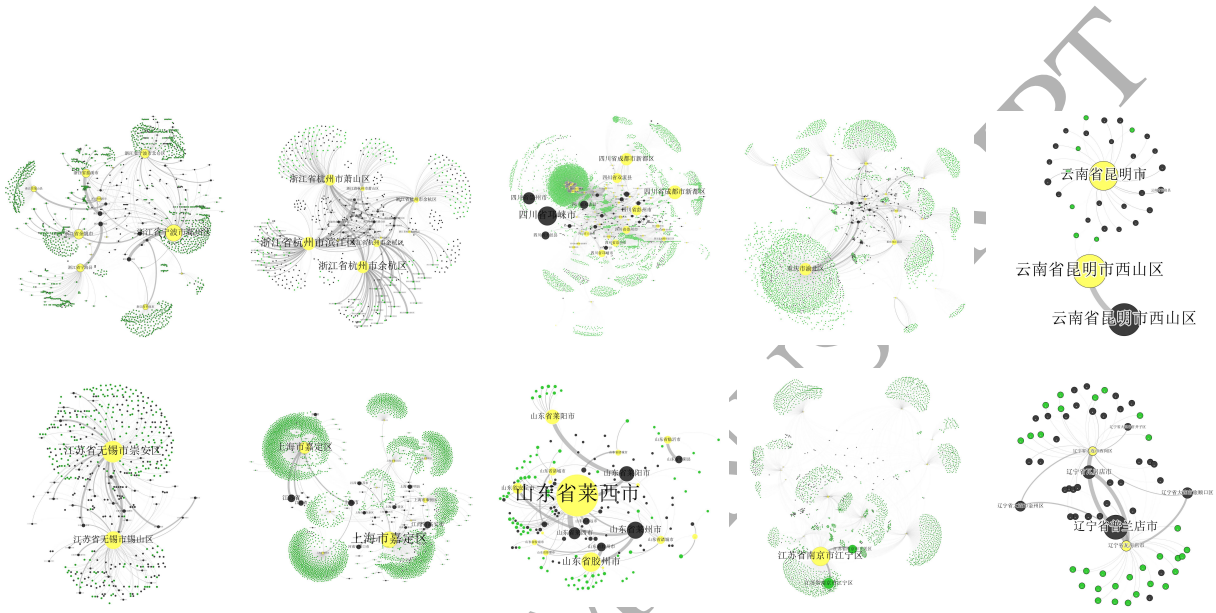

Figure 9: Visualizations of the pork supply chain networks for ten cities in China based on data from NIPTS. Cities from left to right: Ningbo, Hangzhou, Chengdu, Chongqing, and Kunming on the top and Wuxi, Shanghai, Qingdao, Nanjing, and Dalian on the bottom. Colors represent type of node with farms in black, slaughterhouses in yellow, and retailers in green. Node size and edge weights corresponding to the total volume of product flows.

5.2 Network structural analysis

Figure 9 depicts the networks for the 10 pilot cities and in Table 2 we list the size and average in- and out-degree between nodes in each stage. We include the *low*, *medium*, and *high* density networks from Section 3 for comparison, demonstrating that the values chosen for simulation studies are representative of the values in the real data. We also plot in Figure 10 the distribution of the volume of pork carried along all links from farm to slaughterhouse (in yellow) and slaughterhouse to retailer (in green) across the 10 networks. The plot depicts the probability of observing a transaction of volume w', which is normalized by the maximum observed transaction volume for each pair of supply chain stages.

The 10 networks differ widely in size and density of connections. The variation in density is considerable, with some networks exhibiting average densities slightly lower than the *low* density setting explored in the simulation studies (e.g. Qingdao) and other networks exhibiting densities much higher than the *high* density example networks (e.g. Chengdu, Chongqing, Shanghai, and Wuxi). Despite the variability in density, the networks all share a similar connection motif of a proportionally small number of slaughterhouses connected to a much larger number of retailers, visible as hub-and-spoke patterns in the network visualizations. Furthermore, most of the cities feature slaughterhouses supplied by multiple farms and which in turn supply a large number of retailers. With high in- and out-degrees, slaughterhouse nodes act as "bottlenecks" in the forward flow of product through the network. Large bottlenecks are especially visible in networks of the cities Chongqing and Wuxi. We have already seen that density is an important factor in determining traceability. Below we will see that bottlenecks are another important factor in determining (or limiting) traceability.

5.3 Correlation of *NTE* and predictive accuracy in the 10 pilot cities

We start by measuring the correlation between the *NTE* E^λ and the prediction accuracy $\langle\gamma(g)\rangle$ for each network. Following the methodology described in Section 4, we create outbreak scenarios then calculate the probability of accurately identifying the simulated outbreak source, allowing a variable number of guesses $g \in [1, 10]$ for the source farm. Figure 10 presents predictive accuracy at each value of g as a function of *NTE* for outbreaks observed in $\lambda = 1$ retailer (b) and $\lambda = 2$ retailers (c). Results at $g = 1$ guess are indicated in green and at $g = 10$ in yellow.

The results in Figure 10 demonstrate that a clear negative linear relationship exists between E^λ and $\langle\gamma(g)\rangle$ in the pork supply networks. The correlation is more significant for higher values of g. When $g = 10$, the Pearson correlation coefficient between E^1 and $\langle\gamma(1)\rangle$ is -0.92 ($p < 0.001$), while when $g = 1$, the correlation is -0.82 ($p < 0.003$). Similarly, for E^2, the correlation coefficient is -0.88 when $g = 10$ ($p < 0.001$) and -0.64 when $g = 1$ ($p = 0.036$). The observed correlation for higher values of g suggests that when applied to real data, *NTE* is effectively measuring the ability to trace back

Table 2: Summary of network parameters used in stylized networks and observed in pork supply chain data from the NIPTS for ten piloting cities in China.

Stylized Networks	$\|F\|$	$\|D\|$	$\|R\|$	$\bar{d}_F^{out}$	$\bar{d}_{FD}^{in}$	$\bar{d}_{DR}^{out}$	$\bar{d}_{DR}^{in}$
Low	100	20	500	2.0	10.0	30.0	1.2
Medium	100	20	500	3.0	15.0	40.0	1.2
High	100	20	500	4.0	20.0	80.0	3.2
City							
Ningbo (NB)	185	13	1263	1.9	26.5	99.7	1.0
Hangzhou (HZ)	379	9	228	2.2	92.8	25.3	1.0
Chengdu (CD)	312	56	4467	2.8	15.7	585.1	7.3
Chongqing (CQ)	461	14	2827	2.3	76.3	213.3	1.1
Kunming (KM)	25	3	7	1.0	8.3	3.0	1.3
Wuxi (WX)	251	2	222	1.3	157.5	113.0	1.0
Shanghai (SH)	93	10	4472	2.0	18.5	449.4	1.0
Qingdao (QD)	72	15	85	2.5	11.8	5.9	1.1
Nanjing (NJ)	232	12	2350	1.9	36.5	234.1	1.2
Dalian (DN)	37	2	30	1.5	28.5	15.0	1.0

the source of actual outbreaks, whereas for lower values of g the lack of strong correlation indicates that prediction accuracy is not capturing the information about the predictability of accurate source identification as *NTE*.

5.4 Insights for network design with improved traceability

We now discuss how *NTE* might be used to recommend strategies for proactively improving traceability in supply chain networks based on insights gained from the analysis of network structural parameters in the previous sections. One of the core findings from the *Network Traceability Entropy* results is that the scores are widely distributed, suggesting that the ability to identify the source of outbreaks differs markedly across the cities. Values range from a minimum of $\{E^1, E^2\} = \{1.7, 1.4\}$ observed for Qingdao to a maximum of $\{E^1, E^2\} = \{5.6, 5.3\}$ observed for Wuxi. These scores mean that in the event of an outbreak occurring in the pork supply chain in Qingdao, the *effective number* of feasible sources can be narrowed down to $2^{1.4} = 2.6$ farms after two reports of illness, on average. For the same outbreak and illness reporting scenario occurring in Wuxi, the *effective number* of feasible sources is only narrowed down to $2^{5.3} = 39.4$ farms. In practical terms, these figures suggest very different likelihoods of successful investigations to identify the source of an outbreak. Whereas it might be reasonable for public health responders to investigate and sample the 2 to 3 culprit farms supplying pork to Qingdao, the same might not be feasible for the 39 potential culprits supplying

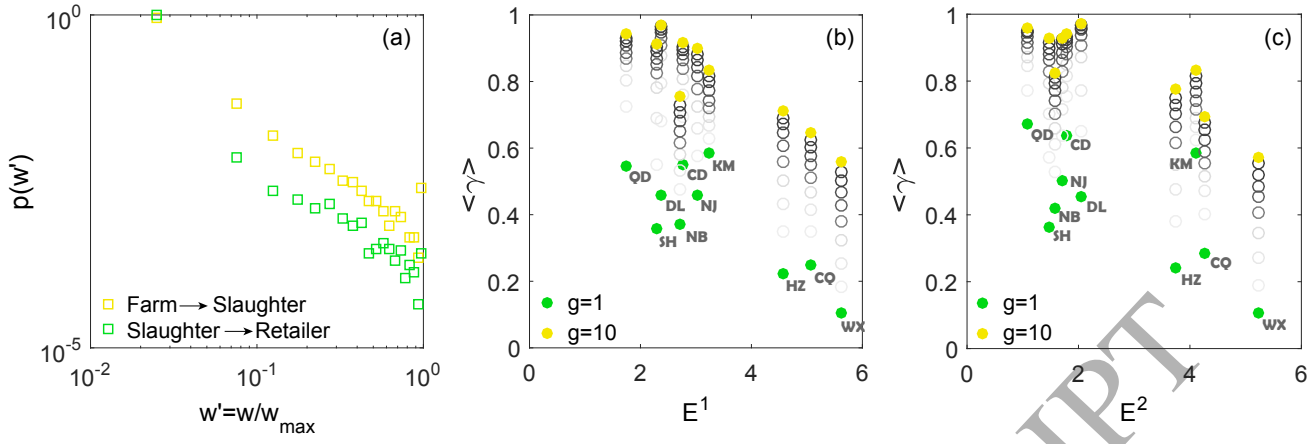

Figure 10: Degree distribution and correlation of traceability entropy and prediction accuracy in pork supply networks for ten cities in China. (a) Distribution of the volume of pork, $w?, carriedalongalllinksfromfarmtoslaughterhouse(inyellow)andslaughterhousetoretailer(ingreen)acros$ is normalized by the maximum transaction volume for each set of supply chain stages. (b) Correlation of traceability entropy and predictive accuracy with one observed illness report and (c) with two observed illness reports. Predictive accuracy results with $g = 1$ guess for the source farm are indicated in green and with $g = 10$ guesses in yellow. Grey circles represent intermediate numbers of guesses.

Wuxi.

For any outbreak occurring in the Wuxi supply chain, the uncertainty in the traceback investigation can be decreased only by waiting for more cases of illness to report, clearly an undesirable solution. However, proactive measures could be taken to improve traceability in supply chains like this by making strategic modifications to the network structure.

For example, we observe that the 3 cities with the highest *NTE* scores, Wuxi, Chongqing, and Hangzhou, exhibit the highest values for average in-degree to the slaughterhouse stage ($\bar{d}^{in}_{FD}$), the parameter indicating bottleneck behavior. To explore the role of this parameter on traceability in the full dataset, in Figure 11 we plot *NTE* as a function of the average in-degree to the slaughterhouse stage. The observed positive correlation suggests that bottlenecks are an important factor in determining traceability. This follows our expectations, since when many farms supply a single slaughterhouse it becomes difficult to distinguish between them at this bottleneck to identify the culprit of an outbreak. This finding suggests that an effective means of improving traceability would be to decrease the number of links into the distribution stage. However, system-level changes such as increasing the number of slaughterhouses or decreasing the number of links into each existing slaughterhouse might not be feasible or even desirable. Strategies might also be imagined to achieve an "effective" increase in the number of slaughterhouses or decrease in the number of links through operational changes. One possible strategy would be to compartmentalize the slaughterhouses such

that each facility is divided into independent sections that process product only from specific farms. In this way, Wuxi's two slaughterhouses (see Table 2) could be divided into 10 sub-sections, each of which trade only with 10% of the farms. The average in-degree to the slaughterhouse stage would become $251/20 = 12.6$, which is comparable to the average in-degree of Qingdao (at 11.8), the supply chain with the lowest entropy score. Of course with any divisions or compartmentalization, it would be important to ensure that essential system flexibility is not lost; this could be verified by designing divisions such that each division is self-sufficient in production and demand of product.

This discussion and any supply chain design changes mentioned are only suggestive. Before implementing any policy or operational changes a full study would be necessary to investigate (i) which structural variables have the greatest influence on traceability, (ii) what combination of parameter values for these variables optimally facilitate traceability, and (iii) what changes could feasibly be implemented to a systemic supply chain network. Nonetheless, the results discussed here present the first steps into the development of a quantified study of how network parameters affect traceability, and consequently, how this knowledge can inform the design or redesign of network structure.

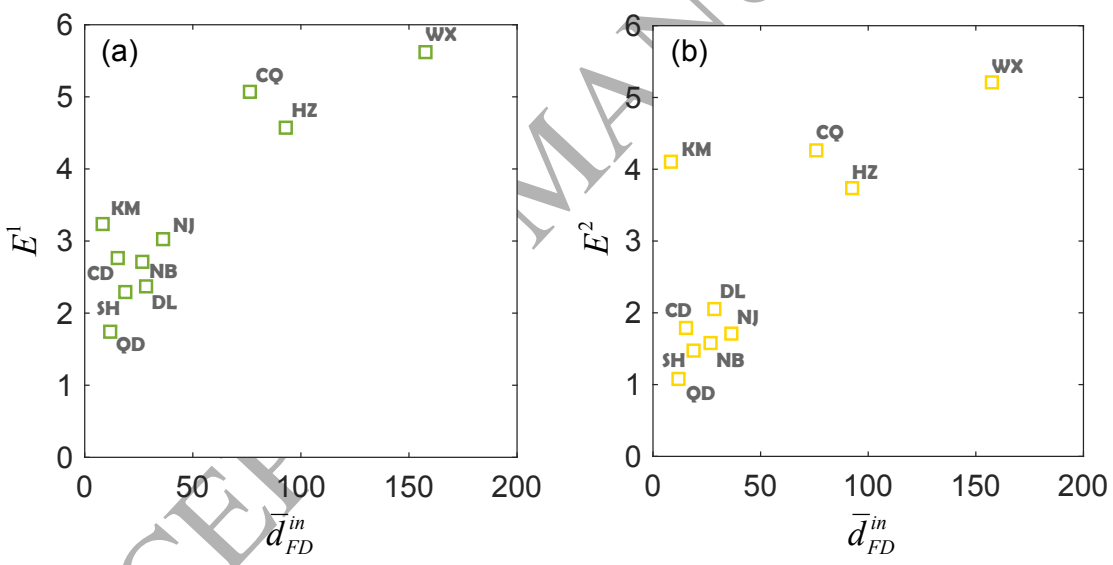

Figure 11: Traceability Entropy as a function of the average in-degree ($\bar{d}^{in}_{FD}$) from farms to slaughterhouses for the 10 case study networks, for (a) one report of illness and (b) two reports of illness.

6 Conclusions and Discussion

While significant work has focused on understanding the role of network structure on propagation dynamics, its impact on traceability, or the ability to identify the propagation source, has received much less attention. In this paper we propose a novel quantity, *Network Traceability Entropy* (*NTE*,

E^{λ}), to measure the ability of a network structure to support accurate source identification. This measure calculates the information-theoretic entropy of the posterior probability distribution for the source location resulting from implementing a network source identification inference algorithm. *NTE* is comprehensive and efficient: by summarizing information about the full posterior probability distribution over feasible sources into a single score, this measure presents an improvement over the existing simulation-based predictive accuracy metric that is based on a single binary outcome or single rank value, while being just as convenient. Illustrative examples have been provided to demonstrate this improvement.

Our work provides the first study systematically evaluating the role of network structural parameters on the ability to identify outbreak sources. Using food supply chain networks as an example and varying a range of topological properties, we use *NTE* as a tool to compare the traceability of various network configurations. This study yields three main observations. First, we have seen that traceability will be higher if a network is less densely connected, or if the degree distributions are more heterogeneous. These results, which demonstrate expected behavior of traceability in response to changes in network structure, serve as an initial validation of *NTE* as an appropriate measure of network traceability.

Second, analysis of the impact on traceability of the amount of "centralization" or "regionalization" in food supply network community structures reveals interesting findings. On the whole, supply networks with high centralization exhibit lower traceability than structures with high regionalization connections. However decreasing the number of cross-regional connections in favor of increasing intra-regional connections will improve traceability only to a point, above which it will drop dramatically. This result suggests there exists an optimal tradeoff, an observation with important implications in the context of ongoing trends in global and local food market development.

Third, we find that the initial volumes of food produced, or the prior distribution over feasible sources, does not bias *NTE* results. This result implies that network structure itself dominates the determination of traceability, highlighting the importance and feasibility of applying this universal measure to quantify, compare, and optimize networks with various topology.

The case study reiterates the results from the stylized network analysis, emphasizing the impact of network density, and in particular, the important role played by bottleneck nodes in limiting traceability. Moreover, the large variability in traceability scores observed across the 10 real supply chain networks studied, indicating accordingly varied likelihoods of successful source identification in the event of an outbreak, highlights relevant information for public health emergency preparedness.

With the powerful ability to diagnose and compare the traceability of various networked settings, this tool has important implications for both policy and managerial decision makers. For example in the context of food distribution, policy decision makers, e.g. the FDA and local health departments, emergency preparedness officials, and other risk assessment bodies in charge of providing public health monitoring, might be interested in proactively computing and comparing the traceabil-

ity scores of the aggregated supply chains for various food products in order to identify high-risk and therefore high-priority items. Highly centralized or highly regionalized supply chains might receive special focus. For supply chains posing a greater risk, resources might be allocated to monitor these items more closely or to insert more preventative controls. For example, the 2011 Food Safety Modernization Act (FSMA) now requires the identification of high-risk foods, monitoring these products with additional recordkeeping requirements [51, 52]. The *NTE* metric could contribute to determining this high-risk product set. Companies with high traceability scores might be exempted from specific requirements, or be rewarded for increasing their traceability score. For managerial decision makers responsible for a company's logistics operations, the traceability metric could be used to diagnose the traceability of existing or proposed supply structures. The results might inform design or redesign of these structures, e.g. diversifying the supply to decrease the number of connections into bottleneck facilities. As suggested in the case study, the traceability score can also be used to determine effective changes to operations that do not require any structural or market changes. In other network and problem contexts, this tool can be used to similarly diagnose high-risk system settings, or to inform the design of network configurations that maximize traceability.

It is important to note that a measure of network traceability is only as good as the network data available to analyze. Access to high-quality network data is therefore the major prerequisite and implementation challenge for any study of traceability. With the rapid expansions of big data technologies such as RFID, Internet-of-Things, and Blockchain, full system network data is becoming both more frequently recorded and more comprehensive. The data from the Chinese "Important Product Traceability System" evaluated in the case study is one such example, with abundant supply data readily available for real-time monitoring and analysis. Still, even the most comprehensive data capture systems will be limited to regulated market data only and cannot make predictions regarding food produced or tampered with illegally. Therefore no study results should be concluded or changes to network structure prescribed without first addressing potential data limitations. For example, with regard to the applications mentioned above, since it is not possible to map any live market supply chain perfectly (as in our case study data), future work would need to be done to identify the sensitivity of the traceability score to missing information on network structure and dynamics.

With these implementation challenges in mind, this work serves as an illustration of what can be learned by applying *NTE* to high-quality network data, when available, and provides a first step into the development of a quantified theory of the relationship between traceability and network structure. The joint effect of various network parameters (e.g., degree distribution, density, flow distribution, community structure, spatial structure) on network traceability opens possibilities for traceability studies from an operational perspective. Future studies should concentrate on identifying combinations of adaptable or flexible features that are not only important in determining traceability, but that also can be modulated in real networks in order to inform the design or remodeling of food supply structures. While a simplified model of outbreak propagation and source estimation

is assumed in this paper, there is an opportunity in the future to augment the analysis here with more comprehensive spatio-temporal probabilistic approaches for contamination source localization, or to augment the model to consider multiple outbreak sources. Future studies may also apply the method in a more realistic setting, such as in combination with electronic track-and-trace information e.g. RFID data [53]. Furthermore, Network Traceability Entropy can be extended to many other network problem settings involving transmission processes, such as identifying the source of defective or counterfeit parts in manufacturing supply chains, disease contagion, virus infection, or rumors spreading in social media.

Author contributions

XL and ALH contributed equally to this work. XL designed the research; ALH designed the source identification model; XL, ALH and JS analyzed the data; XL, ALH, JS and JJ wrote the paper. The authors declare no conflict of interest.

Acknowledgement

The authors gratefully acknowledge Dr. Yu Shiwei, Mr. Ren Xiaotao and Dr. Zhang Nini from the Ministry of Commerce of China (MOFCOM) for supporting the NIPTS data and for policy implication suggestions. We also wish to thank Richard Larson, Simon DeDeo, and Ryan James for helpful discussions. XL is supported by the Natural Science Foundation of China (71522014, 71771213, 71790615) and the Science and Technology Planning Project of Guangdong Province (2015B010131015). During the development of this work ALH was supported by a Robert Wood Johnson Foundation (RWJF) Public Health Services and Systems Research (PHSSR) award, a German Research Foundation (DFG) award, a Bayer Foundation Award, and by the Federal Institute for Risk Assessment (BfR). JS and JJ are supported by the Natural Science Foundation of China (71725001, 71331008, 71731009).

References

[1] T Déirdre Hollingsworth, Neil M Ferguson, and Roy M Anderson. Will travel restrictions control the international spread of pandemic influenza? *Nature Medicine*, 12(5):497–499, May 2006.

[2] Vittoria Colizza, Alain Barrat, Marc Barthelemy, Alain-Jacques Valleron, and Alessandro Vespignani. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLOS Medicine*, 4(1):1–16, 01 2007.

[3] Karima R. Nigmatulina and Richard C. Larson. Living with influenza: Impacts of govern-

ment imposed and voluntarily selected interventions. *European Journal of Operational Research*, 195(2):613–627, 2009.

[4] Stan N. Finkelstein, Richard C. Larson, Karima Nigmatulina, and Anna Teytelman. Engineering effective responses to influenza outbreaks. *Service Science*, 7(2):119–131, 2015.

[5] Vincenzo Fioriti, Marta Chinnici, and Jesus Palomo. Predicting the sources of an outbreak with a spectral technique. *Applied Mathematical Sciences*, 8(135):6775–6782, 2014.

[6] Andrey Y Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E*, 90(1):012801, 2014.

[7] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alejandro Lage-Castellanos, and Riccardo Zecchina. Bayesian inference of epidemics on networks via belief propagation. *Physical Review Letters*, 112(11):118701, March 2014.

[8] Devavrat Shah and Tauhid Zaman. Rumor centrality: A universal source detector. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS 12, pages 199–210, New York, NY, USA, 2012.

[9] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. Finding effectors in social networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1059–1068, New York, NY, USA, 2010.

[10] Cesar Henrique Comin and Luciano da Fontoura Costa. Identifying the starting point of a spreading process in complex networks. *Physical Review E*, 84(5):056105, 2011.

[11] Pedro C. Pinto, Patrick Thiran, and Martin Vetterli. Locating the source of diffusion in large-scale networks. *Physical Review Letters*, 109:068702, Aug 2012.

[12] Dirk Brockmann and Dirk Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342, December 2013.

[13] Abigail L Horn and Hanno Friedrich. Locating the source of large-scale diffusion of foodborne contamination. *arXiv preprint arXiv:1805.03137*, 2018.

[14] F. Morone and H. A. Makse. Influence maximization in complex networks through optimal percolation. *Nature*, 524(7563):65–U122, 2015.

[15] Zi-Ke Zhang, Chuang Liu, Xiu-Xiu Zhan, Xin Lu, Chu-Xu Zhang, and Yi-Cheng Zhang. Dynamics of information diffusion and its applications on complex networks. *Physics Reports*, 651:1–34, 2016.

[16] David L. Alderson. OR Forum - Catching the "network science" bug: Insight and opportunity for the operations researcher. *Operations Research*, 56(5):1047–1065, 2008.

[17] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, 1998.

[18] Cristopher Moore and M. E. J. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61:5678–5682, May 2000.

[19] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86:3200–3203, Apr 2001.

[20] Marc Barthélemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology*, 235(2):275–288, 2005.

[21] Alain Barrat, Marc Barthlemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[22] Daniel Smilkov and Ljupco Kocarev. Influence of the network topology on epidemic spreading. *Physical Review E*, 85(1):016114, 2012.

[23] Christel Kamp, Mathieu Moslonka-Lefebvre, and Samuel Alizon. Epidemic spread on weighted networks. *PLOS Computational Biology*, 9(12):e1003352, 2013.

[24] F. Natale, A. Giovannini, L. Savini, D. Palma, L. Possenti, G. Fiore, and P. Calistri. Network analysis of italian cattle trade patterns and evaluation of risks for potential disease spread. *Preventive veterinary medicine*, 92(4):341–50, 2009.

[25] Paolo Bajardi, Alain Barrat, Fabrizio Natale, Lara Savini, and Vittoria Colizza. Dynamical patterns of cattle trade movements. *PLOS One*, 6(5):e19869, 2011.

[26] Hartmut H. K. Lentz, Thomas Selhorst, and Igor M. Sokolov. Spread of infectious diseases in directed and modular metapopulation networks. *Physical Review E*, 85:066111, 2012.

[27] B. Pinior, U. Platz, U. Ahrens, B. Petersen, F. Conraths, and T. 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(1):25–39, 2012.

[28] Y. Liu and L.M. Wein. Mathematically assessing the consequences of food terrorism scenarios. *Journal of Food Science*, 73(7):346–353, 2008.

[29] Lawrence M. Wein. OR Forum - Homeland security: From mathematical models to policy implementation: The 2008 philip mccord morse lecture. *Operations Research*, 57(4):801–811, 2009.

[30] S. Lazzarini, F. R. Chaddad, and Cook. M. L. Integrating supply chain and network analysis: The study of netchains. *Journal on Chain and Network Science*, 1(1):7–22, 2001.

[31] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[32] SM Dancoff and H Quastler. The information content and error rate of living things. *Essays on the Use of Information Theory in Biology*, 1953.

[33] Hanser Sean F McCowan, Brenda and Laurance R Doyle. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal behaviour*, 57(2):409–419, 1999.

[34] Alfonso Delgado-Bonal and Javier Martín-Torres. Human vision is determined based on information theory. *Scientific reports*, 6:36038, 2016.

[35] Rob Lee, Philip Jonathan, and Pauline Ziman. Pictish symbols revealed as a written language through application of shannon entropy. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 466, pages 2545–2560. The Royal Society, 2010.

[36] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[37] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, July 2012.

[38] Xin Lu, Erik Wetter, Nita Bharti, Andrew J Tatem, and Linus Bengtsson. Approaching the limit of predictability in human mobility. *Scientific Reports*, 3, 2013.

[39] Abigail Lauren Horn. *Locating the source of large scale outbreaks of foodborne disease*. PhD thesis, Massachusetts Institute of Technology, 2016.

[40] Tejas Bhatt, Caitlin Hickey, and Jennifer C. McEntire. Pilot projects for improving product tracing along the food supply system. *Journal of Food Science*, 78(s2):B34–B39, 2013.

[41] Smith Kirk, Miller Ben, Vierk Katie, Williams Ian, and Hedberg Craig. Product tracing in epidemiologic investigations of outbreaks due to commercially distributed food itemsapplication, utility, and considerations, 2015. Accessed June 10, 2017.

[42] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[43] Juliane Manitz, Thomas Kneib, Martin Schlather, Dirk Helbing, and Dirk Brockmann. Origin detection during food-borne disease outbreaks - a case study of the 2011 EHEC/HUS outbreak in Germany. *PLOS Currents Outbreaks*, 6, 04 2014.

[44] Hartmut H. K. Lentz, M. Konschake, K. Teske, M. Kasper, B. Rother, R. Carmanns, B. Petersen, F.J. Conraths, and T. Selhorst. Trade communities and their spatial patterns in the german pork production network. *Preventive Veterinary Medicine*, 98(2-3):176–181, 2011.

[45] Hanno Friedrich. *Simulation of logistics in food retailing for freight transportation analysis*. PhD thesis, Karlsruher Institut für Technologie, 7 2010.

[46] Daniel Grady, Christian Thiemann, and Dirk Brockmann. Robust classification of salient links in complex networks. *Nature Communications*, 3:864, 2012.

[47] Ellen B McCullough, Prabhu L Pingali, and Kostas G Stamoulis. *The Transformation of Agri-Food Systems: Globalization, Supply Chains and Smallholder Farmers*. The Food & Agriculture Organization of the United Nations and Earthscan, London, Sterling, VA, 08 2008.

[48] Robert P King, Michael S Hand, Gigi DiGiacomo, Kate Clancy, Miguel I Gómez, Shermain D Hardesty, Larry Lev, and Edward W McLaughlin. *Comparing the structure, size, and performance of local and mainstream food supply chains*. Economic Research Report, United States Department of Agriculture, Economic Research Service, 06 2010.

[49] David W Hughes and Olga Isengildina-Massa. The economic impact of farmers' markets and a state level locally grown campaign. *Food Policy*, 54:78–84, 2015.

[50] Ministry of Commerce of China, 2017. Accessed June 12, 2017.

[51] US FSMA. HR 2751 FDA food safety modernization act, 2011. Accessed June 10, 2017.

[52] US Congress. HR 3448 public health security and bioterrorism preparedness and response act of 2002, 2002. Accessed June 10, 2017.

[53] A. Regattieri, M. Gamberi, and R. Manzini. Traceability of food products: General framework and experimental evidence. *Journal of Food Engineering*, 81(2):347–356, 2007.