PLOS CURRENTS OUTBREAKS

Origin Detection During Food-borne Disease Outbreaks – A Case Study of the 2011 EHEC/HUS Outbreak in Germany

April 1, 2014 · Research

Juliane Manitz[1], Thomas Kneib[1], Martin Schlather[2], Dirk Helbing[3], Dirk Brockmann[4]

1 Department of Statistics and Econometrics, University of Göttingen, Göttingen, Germany, **2** School of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany, **3** Swiss Federal Institute of Technology, ETH Zurich, Zurich, Switzerland; Risk Center, ETH Zurich, Zurich, Switzerland, **4** Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, United States of America; Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, United States of America; Robert Koch Institute, Berlin, Germany

Manitz J, Kneib T, Schlather M, Helbing D, Brockmann D. Origin Detection During Food-borne Disease Outbreaks – A Case Study of the 2011 EHEC/HUS Outbreak in Germany. PLOS Currents Outbreaks. 2014 Apr 1. Edition 1. doi: 10.1371/currents.outbreaks.f3fdeb08c5b9de7c09ed9cbcef5f01f2.

Abstract

The key challenge during food-borne disease outbreaks, e.g. the 2011 EHEC/HUS outbreak in Germany, is the design of efficient mitigation strategies based on a timely identification of the outbreak's spatial origin. Standard public health procedures typically use case-control studies and tracings along food shipping chains. These methods are time-consuming and suffer from biased data collected slowly in patient interviews. Here we apply a recently developed, network-theoretical method to identify the spatial origin of food-borne disease outbreaks. Thereby, the network captures the transportation routes of contaminated foods. The technique only requires spatial information on case reports regularly collected by public health institutions and a model for the underlying food distribution network. The approach is based on the idea of replacing the conventional geographic distance with an effective distance that is derived from the topological structure of the underlying food distribution network. We show that this approach can efficiently identify most probable epicenters of food-borne disease outbreaks. We assess and discuss the method in the context of the 2011 EHEC epidemic. Based on plausible assumptions on the structure of the national food distribution network, the approach can correctly localize the origin of the 2011 German EHEC/HUS outbreak.

Funding Statement

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) research training group 'Scaling Problems in Statistics' (RTG 1644, www.uni-goettingen.de/en/156579.html) and the Volkswagen Foundation and inspired by FutureICT (www.futurict.eu). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors have declared that no competing interests exist.

Introduction

Due to intensified mass production, facilitated world-wide shipping and novel food manufacturing methods, food-borne disease outbreaks occur more frequently with increasing impacts on society, public health institutions, the economy, and food industry[1]. An estimated 60% of annual gastrointestinal illnesses for each adult in the general population of the United States is caused by food-borne diseases[2]. Moreover, diarrhoea is the second leading cause of morbidity and mortality among children under five years worldwide[3]. food-borne diseases impose enormous financial burden on health care services, routine surveillance and public health investigations, and trigger substantial productivity impacts and product recalls by the food industry. For seven food-borne pathogens an annual burden of $6.5-$34.5 billion in the United States alone was estimated[4].

One of the most substantial challenges in this context is determining the spatial origin of the contaminated food vehicle, which causes the epidemic, for earlier and more effective disease containment. Several factors make detection of the food-borne disease outbreak origin challenging, e.g. population growth, changing eating habits, globalization of food supply chains, production and processing innovations, and microbiological adaptation[1,5]. Furthermore, public health institutes have limited resources to solve issues such as underreporting and low specificity in the association between aetiology and food vehicle[6]. Origin reconstruction is a complex problem because the effects of contaminated food typically occur with a significant time lag and incidence patterns are geographically incoherent. Additionally, specific transport pathways are generally not monitored. More importantly, food distribution networks are multi-scale, spanning length-scale of hundreds to thousands of kilometers, delivering to and within spatially heterogeneous populations. Consequently, it is generically impossible to estimate the geographic origin of the phenomenon based on geometric aspects of the spatial distribution of reported cases. Only for 66% of the outbreaks, public health investigations identified evidence concerning the infection source[7].

These practical difficulties were particularly striking during the German 2011 EHEC (enterohemorrhagic Escherichia coli) outbreak, which affected 3,842 people with unusually high rates of severe HUS (hemolytic-uremic syndrome) cases and mortality. The EHEC/HUS outbreak raised the awareness of timely and efficient origin reconstruction methods and their importance to society, public health institutions, risk assessment authorities and the food industry [2]. There is no general procedure for food-borne disease outbreak investigations, that fits a particular event perfectly. However, the World Health Organization (WHO)[8] provides practical standard guidelines for the investigation and control of food-borne disease outbreaks as a multi-disciplinary task which requires information from many sources.First, an unusual accumulation of disease reports has to be detected and defined as an outbreak. After pathogen specification, initial cases are investigated with regard to common factors and clinical and food specimens are sampled. The corresponding microbiological ‘fingerprinting’ of strains may also identify case relatedness and/or potential sources of contamination.From associated food and environmental samples, backward tracings are initiated to determine the origin. Furthermore, a case definition can be established to identify outbreak related cases and to collect their information on a standardized questionnaire.Using this data, analytical investigations, such as case-control and cohort studies, are performed to test hypotheses about the transmission vehicle and origin. The outbreak source is determined by combining all collected information, otherwise further analytical studies are required.Finally, the potential origin and transmission routes are controlled using forward tracings from contamination to the outbreak cases. Several attempts to improve traceability of food products to their geographical origin have been developed including technical innovations[9], microbiological advances[10], or food forensics[11]. However, detection of outbreak origin remains time-consuming and cost-intensive.

Network theory and network models have become the most important tools for understanding and predicting epidemics in general[12,13,14]. The majority of studies focuses on spatial disease dynamic systems in which networks quantify the coupling strength or transportation fluxes between spatially distributed populations. Almost all studies aim at understanding and forecasting the future time course of an epidemic based on the topological connectivity of the underlying transport networks[15,16]. Furthermore, most studies focus on human-to-human transmissible diseases. Little work has been done, however, on the inverse problem, also known as the ‘zero patient’ problem in epidemics. Shah and Zaman [17,18] developed a universal source detection maximum likelihood estimate, which assumes virus spread in a general graph along a breadth-first-search tree and derive theoretical thresholds for the detection probability. Pinto at al.[19] extended this estimate for partially observed transmission trees. Alternative origin reconstruction methods are based on shortest paths or consequent diameter from transmission trees[20,21]. Prakash et al.[22] and Fioriti and Chinnici[23] developed methods based on spectral techniques to identify a (set of) origin nodes on a transmission network. They utilize a close relationship of source estimation and node centrality as shown by Comin and da Fontoura Costa[24]. However, these methods require comprehensive knowledge of the transmission network, which is rarely the case.

Here we apply a recently developed network-geometric approach for epicenter reconstruction[25] to food-borne diseases. This approach is based on a plausible redefinition of spatial separation and the introduction of an effective distance derived from the underlying food distribution network in combination with viewing the contagion process from the perspective of a specific node in the network. Using the effective distance method, complex spreading patterns can be mapped onto simple, regular wave propagation patterns if and only if the actual outbreak origin is chosen as the reference node. This way, the method can determine the correct outbreak origin based on the degree of regularity of the measured prevalence distribution when viewed in the effective distance perspective. This reconstruction is successful without the knowledge of the detailed infection hierarchy. Here, the underlying network captures the underlying transportation of the contaminated food rather than the mobility pattern of humans.

German EHEC O104:H4/HUS outbreak 2011

Regarding the number of severe HUS cases, the 2011 EHEC/HUS outbreak in Germany, has been the largest E. coli outbreak reported worldwide. Between May 2 and July 26, 2011, 3,842 outbreak associated EHEC cases were reported to the Robert Koch-Institute (RKI), the German Federal Public Health and Surveillance Institute. This included 855 severe HUS cases (22.3%) and 53 patients (1.4%) died. The outbreak was caused by a rare serotype O104:H4 which infected predominantly adults (median age, 43 years), particularly women (68%), and resulted in high HUS and mortality rates[26]. In the previous years, between 925 and 1,283 cases were reported annually, mostly in children. The majority of the infection cases was observed in Northern Germany, which resulted in a higher incidence (number of cases per 100,000 inhabitants) for the corresponding districts than the overall one for Germany (see Fig. 1). Extensive investigations were conducted by the Task Force EHEC, which included a matched case-control study, a recipe-based restaurant cohort study, and backward-/forward-tracings[27]. The entire process was complicated, resource demanding and time-consuming. All investigations required a large amount of data that are typically biased, incomplete, erroneous, and sometimes contradictory. The tracings require a large amount of trained personnel and their success depends on the results of epidemiological studies.Only the combination of several study designs finally lead to the determination of sprouts as the transmission vehicle and the identification of their origin, a farm in Bienenbüttel located in the district Uelzen, Lower Saxony. On June 10, 38 days after outbreak onset, the public was informed to avoid sprout consumption and the responsible production farm was closed.

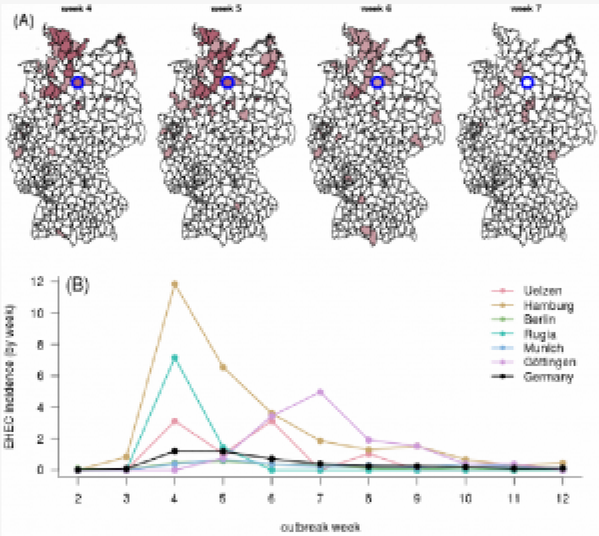

Fig. 1: E. coli incidence in Germany during 2011 EHEC/HUS outbreak.

(A) Each panel depicts a different outbreak week (May 30th until June 20th, 2011). Color intensity quantifies infection counts in for each of the German districts (Data source: [28], Map source: [29]). The alleged origin of outbreak (district Uelzen) is marked in blue. **(B)** Time course of E. coli incidence for selected districts. For reference, the overall German incidence per district is shown in black.

The severe impact of the disease on the population and industry, the fast and wide spread due to mass production and optimized food shipping, and the large public attention emphasize the need for fast and efficient outbreak origin localization.

Network-theoretic origin detection

We consider a model network for spatial food distribution, where nodes $m=1,...,M=412$ represent administrative districts in Germany. Links F_{nm} quantify the amount of goods that are shipped from node m to n per unit time. Note that in the following, we let $F_{nn}=0$. For what follows, only relative flux fractions

$$f_{nm} = \frac{F_{nm}}{\sum_{nm} F_{nm}}$$

(1)

are required to specify the network. The quantities f_{nm} can be interpreted as an effective coupling between districts m and n that is induced by the food distribution between these districts. We consider the quantities f_{nm} as a proxy from which spreading propensities between m and n can be derived.

Because precise measurements of food distribution pathways are not available, we consider an established, approximate heuristic from the social sciences, economics and transportation theory known as the gravity model [30,31]. This approach accounts for the observation that traffic flow increases monotonically with the population size between locations and decreases algebraically with distance, leading to the relationship

$$F_{nm} \propto \frac{N_m^{\alpha} N_n^{\beta}}{(1+d_{nm}/d_0)^{\gamma}},$$

(2)

where N_m, N_n, and d_{nm} quantify the population size of origin m, destination n, and their geographic distance, respectively. The non-negative exponents α, β, γ and distance scale d_0 are parameters of the gravity model [32,33]. Plausible choices for these parameters can be found in the following way: First, we assume that the coupling strength between two locations m and n increase with the number of connections ($N_n \times N_m$) that can be formed between elements of the populations. This implies that $\alpha = \beta$. Additionally, the coupling strength should be proportional to a mean value of the origin and destination population sizes, while leverage by large population nodes should be attenuated. Accounting for this, we choose the geometric average

$$F_{nm} \propto \sqrt{N_n N_m}.$$

(3)

Furthermore, we let the coupling strength F_{nm} decrease with distance. The corresponding tail exponent is consistent with the quantitative assessments of human mobility and transportation networks[34,35], i.e.

$$F_{nm} \propto \frac{1}{d_{nm}^{2+\mu}} \text{with} \, \mu \approx 0.6$$

(4)

Finally, we fix the scale parameter d_0 (in km) in Eq. (2) to be of the order of the average linear extent of a district. With these assumptions, the parameters in the gravity model are $\alpha=\beta=1/2, \gamma=2.6$ and $d_0=10$ km. Although we choose these parameter values as base values, we also investigate the robustness of our results against variations in exponents and found that our results are quite robust.

The gravity model generates a fully connected network with strongly heterogeneous weights, contrasting realistic mobility or transportation networks that possess a sparse topology. In order to obtain a more realistic model for food distribution that exhibits topological sparseness of connections, we follow a procedure recently introduced by Serrano et al. [36]. The idea of this approach is that only links are retained that are statistically significant with respect to a random null model, in which traffic is distributed uniformly among links of a node. Following this idea, we first compute the flux fraction

$$p_{nm} = f_{nm} / \sum_n f_{nm}$$

(5)

for each node m. If at each node, traffic was randomly distributed among the remaining $M-1$ other nodes, a null model would produce $p^0_{nm} \approx 1/M$. Thus, we only retain links that possess a flux fraction larger than $1/M$, i.e. if

$$p_{nm} > 1/M. \tag{6}$$

This approach yields a network skeleton of statistically significant links. Following this procedure the resulting network has an overall connectivity of 18%, see Fig. 2B.

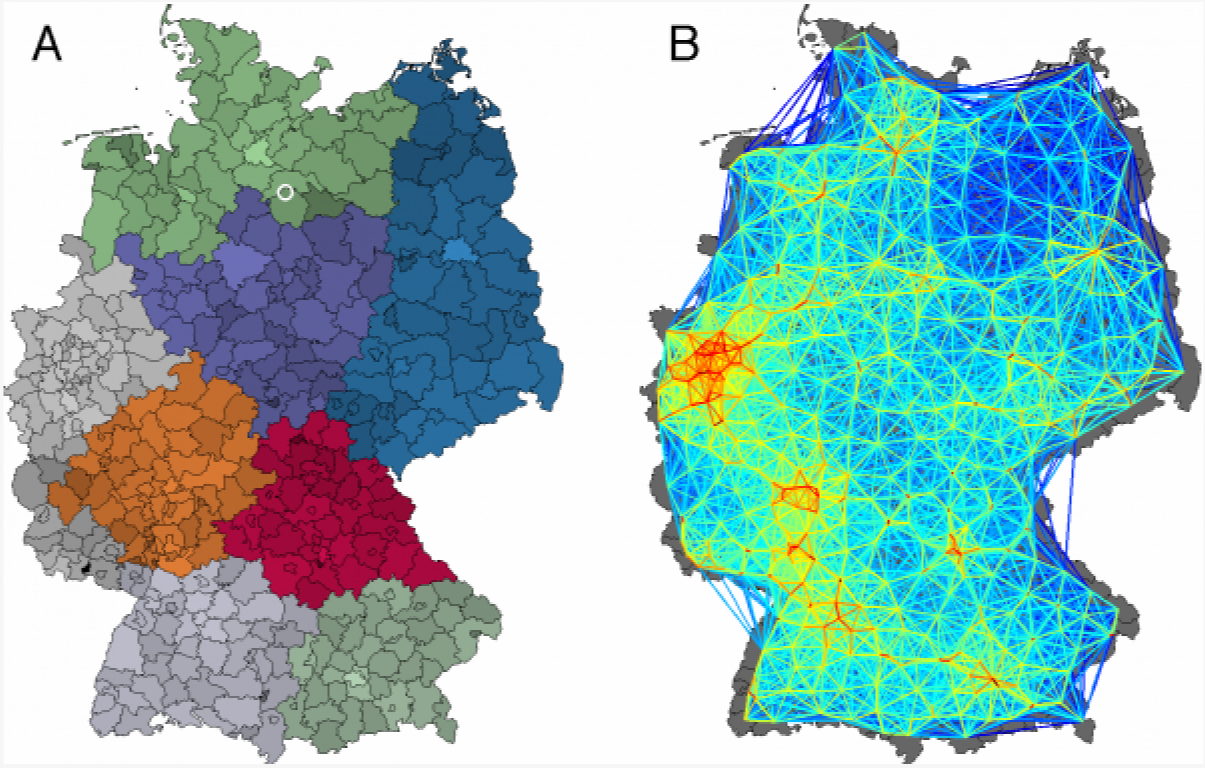

Fig. 2: Multiscale Food Distribution in Germany

(A) A map of German districts; hues correspond to the regional network modules obtained by modularity maximization [37]; color intensity quantifies population density. The origin of the 2011 EHEC/HUS outbreak is marked by a white circle in Bienenbüttel located in the district Uelzen. **(B)** German food shipping network constructed from a gravity model with parameters $\alpha=\beta=1/2, \gamma=2.6$, and $d_0=10$ km. Each district is represented by a network node, coloring corresponds to the link strength. The network has a connectivity of 18.1%.

One of the characteristic features of transportation networks in general, which is also captured by the above gravity model, is its multiscale structure. Although short-range links are usually strongest, the algebraic tail in Eq. (2) yields long-range connections that can dominate spreading phenomena evolving on these networks. Qualitatively, this is illustrated in Fig. 3A which depicts an simple planar quasi-lattice network, in which every node is connected only to its spatially adjacent nodes. Additionally, a few long-range, random connections are added. Because of long-range connections in the network, an initially localized spreading process quickly attains a spatially incoherent structure. As a consequence of this, it is no longer possible to predict with ordinary diffusion when a spreading process will arrive at a given location in the network. More importantly, it is difficult to reconstruct the outbreak origin from a snapshot (or a sequence of snapshots) of the spatio-temporal pattern of spread alone based on conventional planar distance measures and two-dimensional geometry.

Effectively, two nodes that are connected by a long-range link in a multiscale network system are more adjacent than their spatial distance would suggest. Based on this basic and intuitive insight, a recent study[25] introduced the concept of effective distance to network-driven contagion or spreading phenomena. The most important result of this study is that spatio-temporally complex patterns of spreading can be mapped onto simple, regular wave front patterns when conventional distance is replaced by a suitably chosen effective distance. This not only permits calculations of arrival times at any node in the network but, more importantly, the identification of outbreak origins as will be explained in more detail below. The effective distance approach has been shown to work in the context of infectious disease dynamics on a global scale, for instance, the worldwide spread of SARS in 2003 and pandemic influenza H1N1 in 2009.

The effective distance method assumes that, irrespective of the details of the local dynamics of a spreading process, the proliferation of the contagion throughout the network is determined by the coupling between nodes, and that this coupling is quantified by the flux matrix elements f_{nm}. Given an initial outbreak location n_0, a contagion process can take a multitude of paths to any other node in the network. Each path Γ is taken with probability $P(\Gamma)$. Consider a path that starts at n_0 and ends at n_K with a sequence of intermediate steps at nodes $n_i, i=1,...,K-1$ such that

$$\Gamma = \{n_0, . . ., n_K\}.$$

(7)

The probability of the contagion process taking this path is assumed to be given by the product of probabilities of each step

$$P(\Gamma)=\prod_{i=1}^{K}P(n_i|n_{i-1}).$$

(8)

Here, for every link in the network the function $P(n|m)$ is the probability that a contaminated food at m is moved to n. The fundamental assumption in Brockmann and Helbing **25** is that the single step probability $P(n|m)$ is identified with the flux fraction p_{nm} that is determined by the underlying transportation network:

$$P(n|m) = p_{nm} = \frac{f_{nm}}{\sum_n f_{nm}}.$$

(9)

Then, we define the effective distance of a multi-leg path Γ by

$$\Lambda(\Gamma) = K - \log[P(\Gamma)],$$

(10)

where K is the number of links composing the path and $P(\Gamma)$ the corresponding path probability. For the sake of motivation and interpretation, we can decompose the path length into contributions by direct links of this formula:

$$\Lambda(\Gamma) = \sum_{i=1}^{K} \lambda(n_i | n_{i-1}).$$

(11)

Here, the effective length of a direct link $n_{i-1} \to n_i$ is given by

$$\lambda(n_i|n_{i-1}) = 1 - \log P(n_i|n_{i-1}).$$

(12)

This relation establishes a connection between network topological features and effective distance. The functional form is chosen such that a number of important features are fulfilled: (i) the length from n_{i-1} to n_i decreases with increasing probability $P(n_i|n_{i-1})$. That is, for large values of $P(n_i|n_{i-1})$, the effective length is small and for vanishing transition probability the effective length diverges. (ii) The effective length of a multi-step path Γ as defined in Eq. (7) is the sum of the effective lengths of each segment in the path. (iii) Given two paths that occur with certainty (e.g. with $P(n_i|n_{i-1})=1$ for each link), but have a different number of segments, the path that has more segments also has a larger effective length.

Generically, transportation networks are strongly heterogeneous such that, in an ensemble of paths with origin n_0 and destination n_K, the dynamics are dominated by the most probable path and therefore the path of minimum effective length **25**. The effective distance $D(n|m)$ is defined as the minimum effective length of a path $\Lambda(\Gamma)$ from origin m to destination n:

$$D(n|m) = \min_{\Gamma} \Lambda(\Gamma).$$

(13)

From the perspective of a chosen root or reference node m, one can compute the shortest path tree T_m, which is the collection of shortest effective paths to all other nodes in the network. This shortest path tree is equivalent to the most probable contagion hierarchy that a spreading process will take through the network.

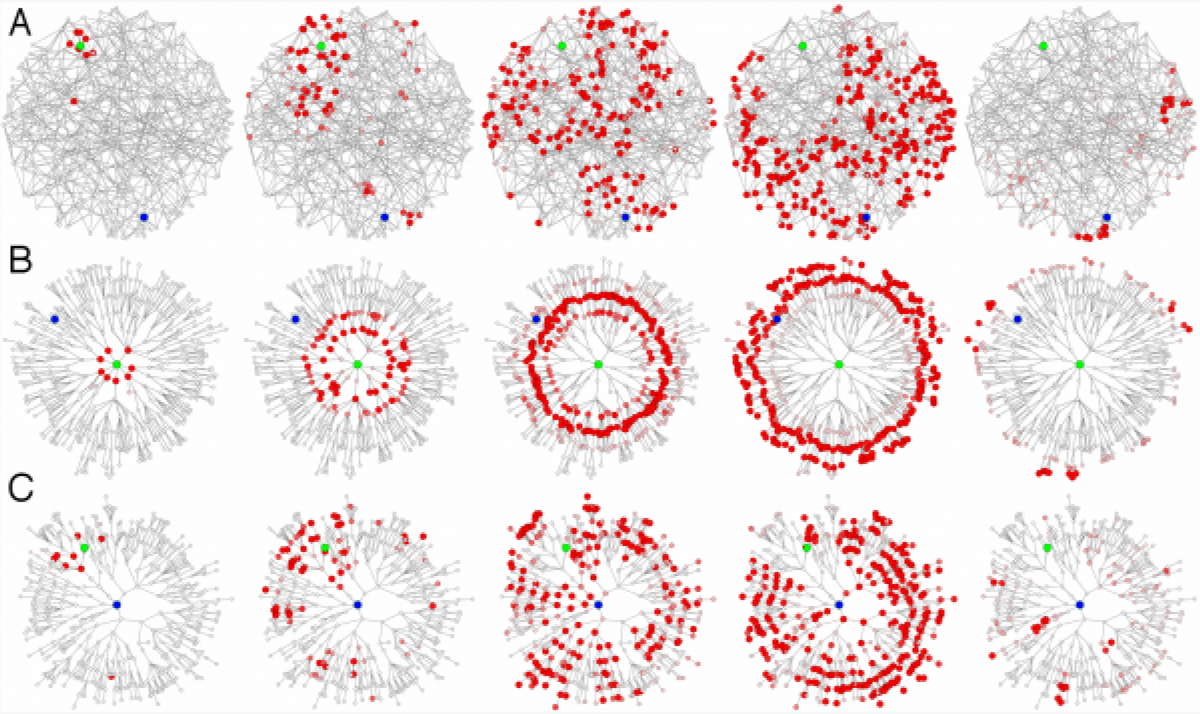

Fig. 3: Effective distance and outbreak origin reconstruction in multi-scale network contagion processes.

(A) Each panel depicts a temporal snapshot (from left to right at equidistant time intervals) in a simple contagion process in which infected nodes (red) deliver the infection to connected nodes at a fixed rate before they recover at a another rate (SIR dynamics [38]). The network consists of 512 nodes on a quasi-triangular, random lattice. Each node is connected to its nearest local neighbors. In addition to the local lattice structure, 128 long range links exist between randomly chosen pairs of nodes. The origin of the outbreak is marked in green. Because of long range connectivity the pattern quickly loses spatial structure and becomes chaotic such that it is difficult to predict from metric cues alone when the contagion arrives at a given node. More importantly, long range connectivity leads to a loss of spatial coherence and it becomes impossible to determine the origin of outbreak. **(B)** The same pattern as in (A) is shown in the effective distance perspective from the outbreak origin. The depicted tree is the shortest path tree, i.e. the most probable spreading path of the contagion process. Radial distance is proportional to effective distance as defined in the text. In this alternative representation the complex pattern in the conventional view is mapped onto a simple propagating wave front and arrival times are easily computed. **(C)** The regularity of the pattern is only present from the perspective of the actual outbreak origin. When the contagion process is viewed from any other node (here the node depicted in blue), the pattern lacks regularity.

Fig. 3B illustrates the advantages of this approach in an artificial multi-scale network. From the perspective of the outbreak origin, the shortest path tree of the root node is shown, and the radial distance in the new map corresponds to the effective distance from the root node to the remaining nodes in the network. The same spreading process that appears to be spatio-temporally complex in the conventional metric layout is equivalent to a regular, constant-speed spreading wave in the effective distance representation. Consequently, one can calculate arrival times based on effective distance alone. In fact, in Brockmann and Helbing[25] it was shown that effective distance from the outbreak origin and arrival time strongly correlate in real scenarios, e.g. the 2003 SARS epidemic and the 2009 H1N1 pandemic influenza outbreak.

The most relevant consequence of the effective distance approach is that, only from the perspective of the actual outbreak origin, the pattern exhibits a regular concentric wave front structure. From the perspective of any other node in the network, the pattern exhibits a more or less disordered structure. Fig. 3C illustrates this. The panels depict the same dynamics as in the other panels from a randomly chosen reference node. Clearly, any spatial regularity is absent. One can now make use of this observation, i.e. the fact that the spreading pattern is regular only from the perspective of the actual outbreak location, to reconstruct the outbreak origin. Given a snapshot of the disease spread, e.g. the disease incidence at every node, one computes the effective distance perspective for each node in the network and quantifies, from which node the pattern appears to be most regular. The node with maximum regularity is considered to be the most likely outbreak origin. In the following we apply this approach to the 2011 EHEC/HUS outbreak in Germany.

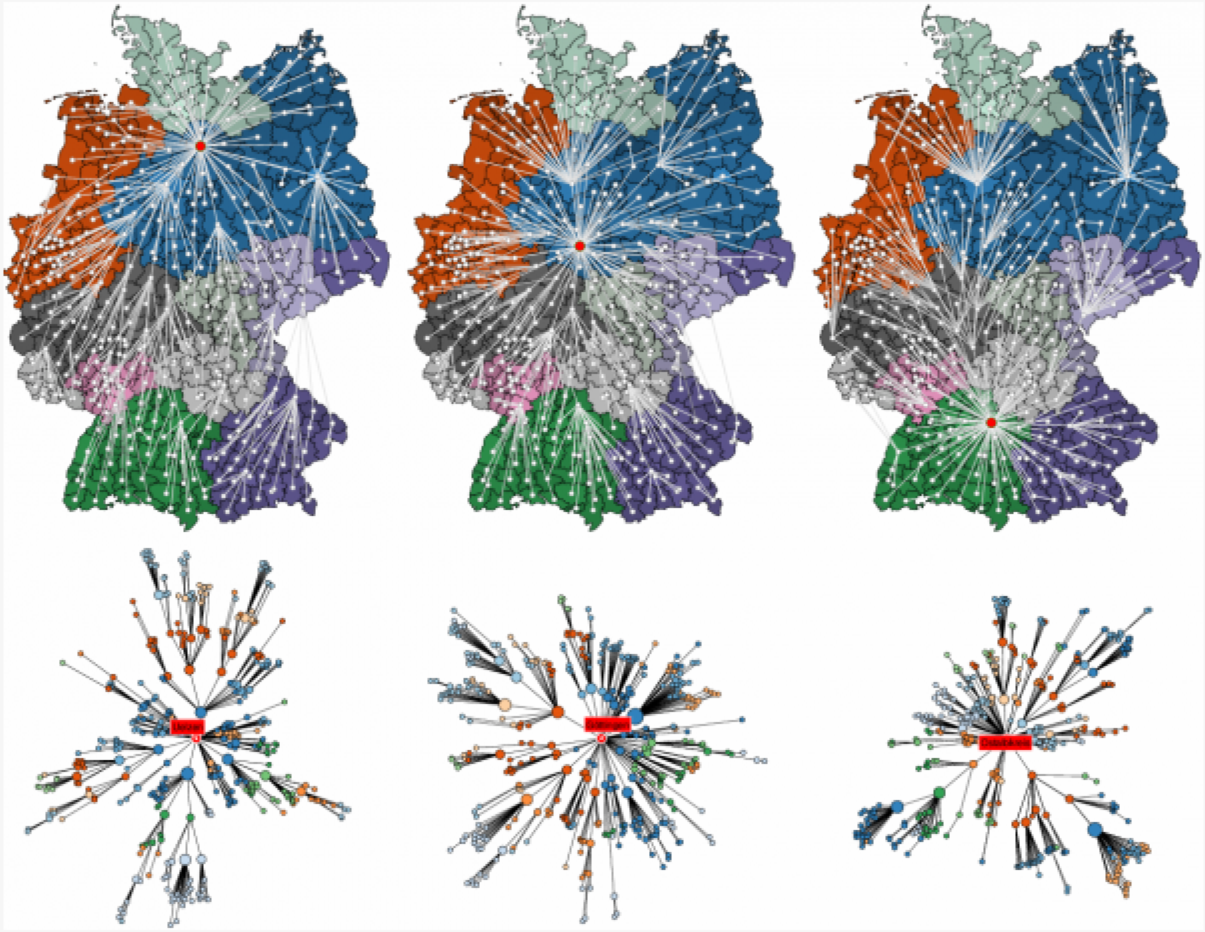

Fig. 4: Shortest path trees and effective distance among districts in Germany.

Each column depicts the shortest path tree T_m for a sample root node (red), from left to right districts Uelzen, Göttingen, and Oberalbkreis. The top row depicts embedded in the conventional geographic representation, the bottom illustrates the shortest path tree in a layout such that the radial distance is proportional to the effective distance from the root node in the same way as in Fig. 3. The shortest path tree T_m represents the most probable path that a contagion process takes with initial outbreak in node m.

Detection of the German EHEC/HUS outbreak origin

Given the gravity model network for food transportation, we first compute the shortest path tree T_m for every potential root node m, see Fig. 4 for examples. Next, a temporal snapshot of the EHEC incidence pattern is analyzed in each of the shortest path tree representations, i.e. from the perspective of all network nodes as potential candidate origins of outbreak. The incidence pattern typically consists of a subset Ω of nodes with non-zero incidence. From the perspective of the actual outbreak origin, the effective distance to these affected nodes, should be small and exhibit a small variance, a consequence of the concentricity of the spreading pattern in the effective distance representation. Therefore, in order to quantify the regularity of the incidence pattern from every potential outbreak origin, we compute the average $\mu(D;m)$ and standard deviation $\sigma(D;m)$ of effective distances to nodes with nonzero incidence (the subset of nodes Ω) **25**.

$$\mu(D;m)=\frac{1}{N_\Omega}\sum_{n\in\Omega}D(n|m),\ \sigma^2(D;m)=\frac{1}{N_\Omega}\sum_{n\in\Omega}D(n|m)^2-\mu(D;m)^2$$

(14)

In combination, small mean and variance are equivalent to high concentricity and, thus, high likelihood that the chosen reference node is the likely outbreak origin.

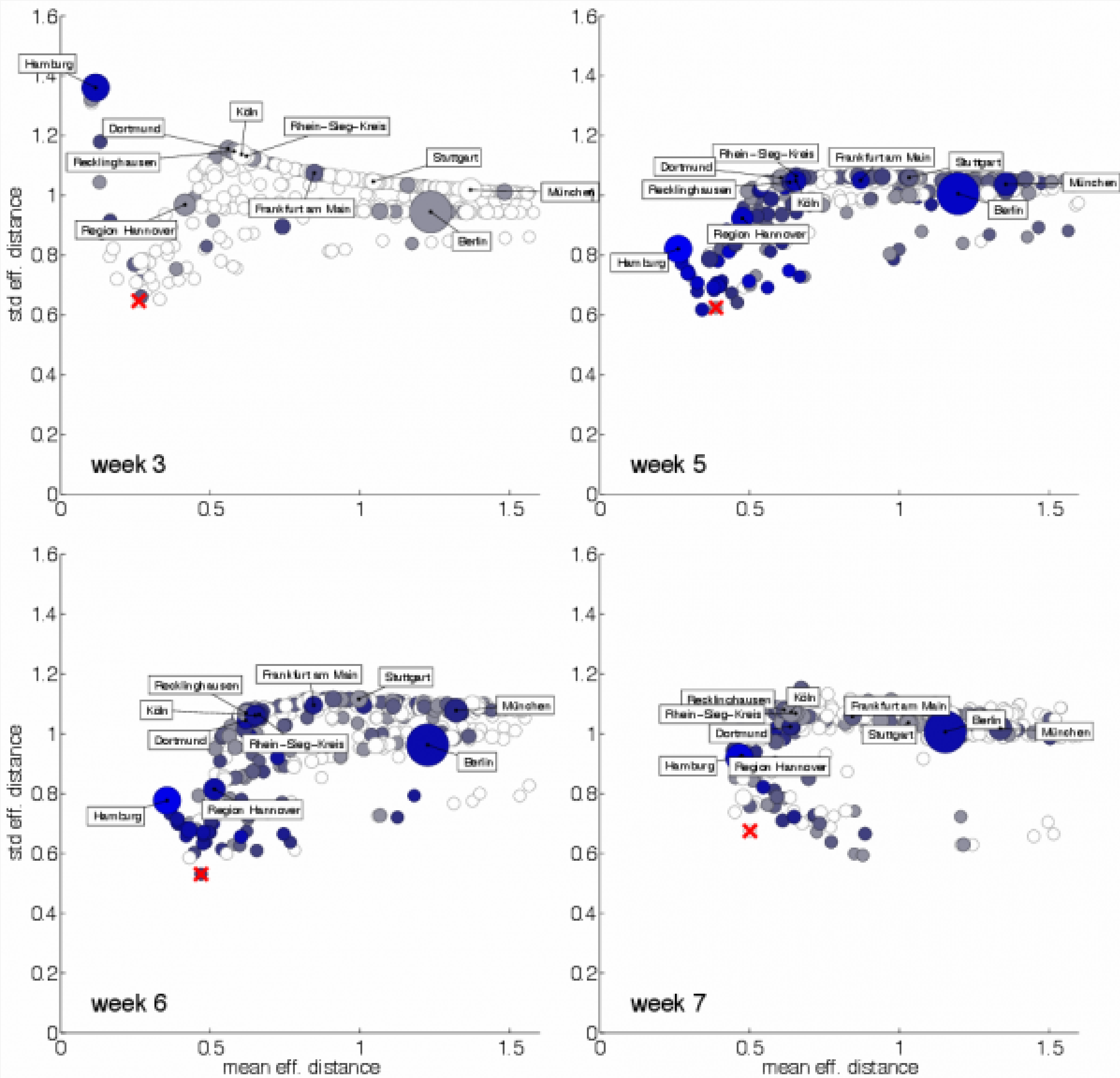

Fig. 5: EHEC/HUS outbreak origin reconstruction

Each panel depicts a scatterplot of mean $\mu(D,m)$ and standard deviation $\sigma(D,m)$ (see Eqs. (14)) of effective distances from candidate nodes m to the subset Ω of nodes that have nonzero incidence for weeks $t=3,5,6,7$ after outbreak onset. All districts are considered as potential candidates as outbreak origin. Symbol size quantifies population size of each district, blueness quantifies incidence in the respective week. A few large district are labeled. The district with combined minimal mean and variance (closest to the origin) has a high likelihood of being the actual 2011 EHEC/HUS outbreak origin. The actual outbreak origin Uelzen in marked by a red cross.

We used the public available E. coli case count data with report date between calendar weeks 18 and 26 of 2011 [28]. According to the Task Force EHEC, this corresponds to the entire outbreak duration from May 2nd until July 4th, 2011 [26]. Fig. 5 shows the results of origin detection when the effective distance approach in combination with a gravity model for food distribution is applied to the EHEC incidence data. Since an E. coli infection clustering was noticed at May 19th, 2011 (outbreak week 3), we computed the mean $\mu(D,m)$ and standard deviation $\sigma(D,m)$ pair for weeks $t=3,5,6,7$ and every node in the network treating every node m as a potential outbreak origin. When both quantities are small, the resulting spreading patterns is most concentric in the effective distance perspective. Fig. 5 shows that already in week 3 of the event, district Uelzen is identified as the potential origin of the outbreak, this is also true for weeks 6 and 7. In week 5 the method incorrectly identifies district Lüneburg as the likely outbreak origin and Uelzen ranks third in the epicenter reconstruction. Note that the geographic center of district Lüneburg is as close to Bienenbüttel (the alleged location of contaminated sprouts) as the geographic center of Uelzen (ca. 20km). Note also, that the overall distribution of pairs $(\mu(D,m),\sigma(D,m))$ differs considerably for each temporal snapshot of EHEC incidence districts close to the actual outbreak location exhibit combined small values of $(\mu(D,m),\sigma(D,m))$. Table in Fig. 6 ranks the candidate outbreak locations for weeks 2 to 8. The ranks were computed by comparing the effective distance to the origin in the $(\mu(D,m),\sigma(D,m))$ scatter plot. For all time windows except weeks 4 and 8 the correct district ranks among the top candidates for EHEC outbreak origin. Note that other potential outbreak origins are typically districts that are in close geographic proximity to the actual outbreak location. This implies that even if the origin cannot be identified on the scale of a single district, potential candidates according to the effective distance methods are confined to a small region in the vicinity of the actual outbreak location, for instance the set of neighboring districts.

Rank	week 3	week 4	week 5
1	Uelzen (5.2 km)	Segeberg (65.8 km)	Lüneburg (5.2 km)
2	Lüneburg (5.2 km)	Harburg (17.9 km)	Steinburg (88.3 km)
3	Cuxhaven (97.9 km)	Steinburg (88.3 km)	Uelzen (5.2 km)
4	Steinburg (88.3 km)	Stade (57.7 km)	Neumünster (103.6 km)
5	Ostholstein (83.0 km)	Herzogtum Lauenburg (25.3 km)	Lübeck (70.7 km)
6	Bremerhaven (128.8 km)	Pinneberg (66.0 km)	Segeberg (65.8 km)
7	Dithmarschen (155.6 km)	Stormarn (43.0 km)	Stade (57.7 km)
8	Lübeck (70.7 km)	Lüneburg (5.2 km)	Herzogtum Lauenburg (25.3 km)
9	Nordwestmecklenburg (109.6 km)	Hamburg (33.0 km)	Harburg (17.9 km)
10	Stade (57.7 km)	Neumünster (103.6 km)	Ostholstein (83.0 km)

Rank	week 6	week 7	week 8
1	Uelzen (5.2 km)	Uelzen (5.2 km)	Bremen (100.3 km)
2	Lüneburg (5.2 km)	Soltau-Fallingbostel (19.9 km)	Delmenhorst (138.8 km)
3	Soltau-Fallingbostel (19.9 km)	Lüneburg (5.2 km)	Osterholz (94.4 km)
4	Steinburg (88.3 km)	Bremen (100.3 km)	Verden (70.8 km)
5	Herzogtum Lauenburg (25.3 km)	Delmenhorst (138.8 km)	Uelzen (5.2 km)
6	Stade (57.7 km)	Verden (70.8 km)	Soltau-Fallingbostel (19.9 km)
7	Harburg (17.9 km)	Osterholz (94.4 km)	Oldenburg (122.2 km)
8	Lübeck (70.7 km)	Oldenburg (122.2 km)	Bremerhaven (128.8 km)
9	Segeberg (65.8 km)	Celle (28.1 km)	Cuxhaven (97.9 km)
10	Bremen (100.3 km)	Cuxhaven (97.9 km)	Oldenburg (Oldenburg) (140.7 km)

Rank	week 9
1	Lüneburg (5.2 km)
2	Uelzen (5.2 km)
3	Stade (57.7 km)
4	Neumünster (103.6 km)
5	Steinburg (88.3 km)
6	Lübeck (70.7 km)
7	Pinneberg (66.0 km)
8	Segeberg (65.8 km)
9	Herzogtum Lauenburg (25.3 km)
10	Cuxhaven (97.9 km)

Fig. 6: EHEC/HUS outbreak origin reconstruction

For each week 2-9 relative to the beginning of the EHEC/HUS outbreak and for each node m in the network a rank was computed based on minimization of a concentricity score $q(m)=\sqrt{\mu 2(D,m)+\sigma 2(D,m)}$. District Uelzen, the actual outbreak district is robustly ranked among the top ranked districts, in weeks 3, 6 and 7, Uelzen is ranked first. We considered all 412 districts. For each district the distance provided in parenthesis represents the approximate distance to the actual outbreak location Bienenbüttel in district Uelzen.

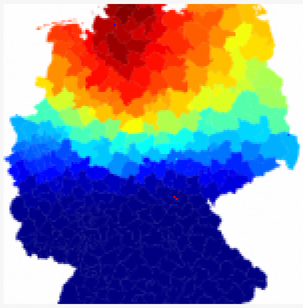

Fig. 7: Correlation of effective distance and arrival time during the German EHEC/HUS outbreak, 2011.

For each district as a potential outbreak origin, we computed the correlation coefficient of arrival time $T(n)$ at every other node n and effective distance from m to n. The magnitude of the correlation coefficient is color-coded from blue to red, corresponding to low and high correlation, respectively. High correlation, corresponding to high likelihood of being the outbreak origin is observed in a spatially coherent region in Northern Germany.

The effective distance method provides an alternative method for outbreak origin reconstruction. An important result presented in Ref. [25] is that arrival times of a network-driven contagion process correlate strongly with effective distance. In fact, the arrival time T_n of the process at a node n with initial outbreak at node m increases linearly with effective distance $D(n|m)$. Again, arrival time and effective distance only correlate strongly when the actual outbreak origin is chosen as the reference node. To supplement the above analysis we computed the correlation coefficient $c(m)$ of arrival times (i.e. the week of reported first case of EHEC/HUS in a given district) with effective distance, considering each node m of the 412 districts as the potential outbreak origin. We then ranked these correlation coefficients. Fig. 7 depicts the magnitude of $c(m)$ in a map of all German districts. Clearly, this method identifies a well-defined region in Northern Germany as containing the likely outbreak location. Note that, in contrast to the incidence patterns, the correlation coefficient varies smoothly with distance from the epicenter somewhere in Northern Germany. When correlation coefficients are ranked according to magnitude, the correct origin district Uelzen only ranks 30 out of 412 districts. However, the difference in correlation coefficients is small among the top-ranked districts, see Table in Fig. 8. The reason for the comparatively low performance of the correlation-based outbreak reconstruction could be that the temporal resolution of the data is too coarse and fluctuations dominate the signal. For instance, travel-related cases could warp the infection pattern. We conclude that outbreak origin reconstruction based on the topological features of the wave front in effective distance, as presented in Fig. 5 and Table in Fig. 6, is a more reliable technique for the detection of the outbreak origin than the correlation approach. Also, for the topological rather than the correlation-based approach only single temporal snapshots of incidence are required, which is an additional advantage.

Rank	District	Corr.	Rank	District	Corr.	Rank	District	Corr.
1	Rendsburg-Eckernförde	0.4848	51	Parchim	0.4132	101	Barnim	0.3443
2	Steinburg	0.4836	52	Güstrow	0.4132	102	Harz	0.3421
3	Segeberg	0.4805	53	Prignitz	0.4090	103	Märkisch-Oderland	0.3405
4	Rotenburg (Wümme)	0.4581	54	Vechta	0.4086	104	Salzlandkreis	0.3373
5	Dithmarschen	0.4569	55	Schaumburg	0.4079	105	Oder-Spree	0.3361
6	Harburg	0.4563	56	Wolfsburg	0.4078	106	Frankfurt (Oder)	0.3337
7	Neumünster	0.4543	57	Peine	0.4056	107	Osterode am Harz	0.3329
8	Stade	0.4542	58	Minden-Lübbecke	0.4051	108	Warendorf	0.3325
9	Stormarn	0.4535	59	Hameln-Pyrmont	0.4041	109	Münster	0.3307
10	Pinneberg	0.4531	60	Hildesheim	0.4041	110	Höxter	0.3256
11	Cuxhaven	0.4531	61	Bad Doberan	0.4036	111	Paderborn	0.3126
12	Kiel	0.4524	62	Rostock	0.4035	112	Cottbus	0.3119
13	Bremerhaven	0.4518	63	Braunschweig	0.4031	113	Spree-Neiße	0.3095
14	Bremen	0.4489	64	Müritz	0.4007	114	Dahme-Spreewald	0.3085
15	Verden	0.4488	65	Nordvorpommern	0.3997	115	Teltow-Fläming	0.3077
16	Soltau-Fallingbostel	0.4469	66	Emsland	0.3996	116	Dessau-Roßlau	0.3067
17	Osterholz	0.4465	67	Rügen	0.3976	117	Wittenberg	0.3054
18	Plön	0.4465	68	Stralsund	0.3962	118	Nordhausen	0.3054
19	Hamburg	0.4460	69	Stendal	0.3958	119	Soest	0.3021
20	Schleswig-Flensburg	0.4455	70	Osnabrück	0.3939	120	Borken	0.3013
21	Flensburg	0.4447	71	Herford	0.3938	121	Coesfeld	0.2985
22	Herzogtum Lauenburg	0.4427	72	Helmstedt	0.3913	122	Anhalt-Bitterfeld	0.2974
23	Lübeck	0.4421	73	Salzgitter	0.3913	123	Göttingen	0.2960
24	Ostholstein	0.4408	74	Demmin	0.3898	124	Kassel	0.2834
25	Nordfriesland	0.4402	75	Greifswald	0.3897	125	Mansfeld-Südharz	0.2817
26	Delmenhorst	0.4402	76	Osnabrück	0.3866	126	Hamm	0.2814
27	Lüneburg	0.4376	77	Ostprignitz-Ruppin	0.3864	127	Eichsfeld	0.2757
28	Wesermarsch	0.4372	78	Wolfenbüttel	0.3862	128	Kyffhäuserkreis	0.2713
29	Diepholz	0.4327	79	Ostvorpommern	0.3849	129	Kassel	0.2674
30	Uelzen	0.4327	80	Lippe	0.3846	130	Hochsauerlandkreis	0.2570
31	Celle	0.4323	81	Uecker-Randow	0.3836	131	Oberspreewald-Lausitz	0.2549
32	Region Hannover	0.4309	82	Börde	0.3810	132	Unna	0.2540
33	Oldenburg	0.4291	83	Holzminden	0.3808	133	Elbe-Elster	0.2538
34	Nienburg (Weser)	0.4283	84	Havelland	0.3778	134	Halle (Saale)	0.2522
35	Nordwestmecklenburg	0.4282	85	Goslar	0.3776	135	Saalekreis	0.2517
36	Oldenburg (Oldenburg)	0.4276	86	Uckermark	0.3768	136	Waldeck-Frankenberg	0.2463
37	Wittmund	0.4275	87	Grafschaft Bentheim	0.3756	137	Bottrop	0.2452
38	Wilhelmshaven	0.4264	88	Brandenburg an der Havel	0.3698	138	Recklinghausen	0.2449
39	Ludwigslust	0.4254	89	Bielefeld	0.3695	139	Nordsachsen	0.2443
40	Friesland	0.4247	90	Neubrandenburg	0.3695	140	Unstrut-Hainich-Kreis	0.2423
41	Lüchow-Dannenberg	0.4246	91	Northeim	0.3650	141	Werra-Meißner-Kreis	0.2408
42	Ammerland	0.4244	92	Magdeburg	0.3637	142	Görlitz	0.2370
43	Aurich	0.4218	93	Mecklenburg-Strelitz	0.3633	143	Herne	0.2347
44	Schwerin	0.4218	94	Oberhavel	0.3596	144	Wesel	0.2343
45	Wismar	0.4212	95	Jerichower Land	0.3583	145	Kleve	0.2336
46	Gifhorn	0.4202	96	Potsdam-Mittelmark	0.3582	146	Dortmund	0.2334
47	Leer	0.4182	97	Gütersloh	0.3533	147	Bautzen	0.2327
48	Cloppenburg	0.4168	98	Steinfurt	0.3532	148	Gelsenkirchen	0.2282
49	Altmarkkreis Salzwedel	0.4157	99	Berlin	0.3516	149	Oberhausen	0.2228
50	Emden	0.4143	100	Potsdam	0.3457	150	Leipzig	0.2217

Fig. 8: Effective distance and arrival time analysis

For each potential district m as outbreak origin we computed the Pearson correlation of arrival time $T(n)$ and effective distance $D(n|m)$ and ranked all districts with respect to correlation magnitude. The actual outbreak origin Uelzen is ranked at position 30. High correlation districts all lie in Northern Germany.

Discussion and conclusion

We introduced a fast and efficient approach for the identification of the origin during food-borne disease outbreaks and evaluated the approach in the context of the 2011 EHEC/HUS outbreak in Germany. A clear advantage of the method is the robust performance on the basis of limited case report data and plausible topological assumptions concerning the underlying food distribution network. When applied to the 2011 EHEC/HUS outbreak in Germany, our method was able to identify an outbreak origin in close proximity to the actual outbreak location (Uelzen, Lower Saxony). Already three days (May 22nd, 2011) after spatial infection clustering, the effective distance approach was able to reconstruct the actual origin. This is particularly promising, as in the context of EHEC/HUS, conventional outbreak investigations, including case-control- and cohort-studies as well as sample testings and tracings along the food-shipping chain,wrongly suggested tomatoes, leafy salads and cucumbers as contaminated foods. When specific suspicions arose that cucumbers imported in Hamburg would be the infection source, our method classifies Hamburg to be a very unlikely origin. The consideration of such contradictory information could have lead to more spatially targeted sample testing, and, therefore could have improved the efficiency of the outbreak investigations.

We believe that this method can complement conventional methods of origin localization of food-borne diseases and consequently facilitate a more timely success which is vital for the development of containment strategies. The underlying network definition by the gravity model is very flexible, so that the transmission vehicle does not has to be known. Basically, the network could also capture a combination of food transportation routes as well as human mobility pattern. As our method is structurally quite general and just derived from topological features of the underlying distribution networks, we believe that our approach may be adapted and applied to a variety of contagion phenomena, human-to-human transmissible diseases, and disease dynamics on individual based contact networks and human-mediated bioinvasion processes.

References

1. Newell DG, Koopmans M, Verhoef L, Duizer E, Aidara-Kane A, Sprong H, Opsteegh M, Langelaar M, Threfall J, Scheutz F, van der Giessen J, Kruse H. Food-borne diseases - the challenges of 20 years ago still persist while new ones continue to emerge. Int J Food Microbiol. 2010 May 30;139 Suppl 1:S3-15. PubMed PMID:20153070.

2. Jones TF, McMillian MB, Scallan E, Frenzen PD, Cronquist AB, Thomas S, Angulo FJ. A population-based estimate of the substantial burden of diarrhoeal disease in the United States; FoodNet, 1996-2003. Epidemiol Infect. 2007 Feb;135(2):293-301. PubMed PMID:17291364.

3. Bryce J, Boschi-Pinto C, Shibuya K, Black RE, the WHO Child Health Epidemiology Reference Group. WHO estimates of the causes of death in children; The Lancet, 26 March–1 April 2005; 365(9465):1147-1152.

4. Buzby JC, Roberts T. Economic costs and trade impacts of microbial foodborne illness. World Health Stat Q. 1997;50(1-2):57-66. PubMed PMID:9282387.

5. Altekruse SF, Cohen ML, Swerdlow DL. Emerging foodborne diseases. Emerg Infect Dis. 1997 Jul-Sep;3(3):285-93. PubMed PMID:9284372.

6. Greig JD, Ravel A. Analysis of foodborne outbreak data reported internationally for source attribution. Int J Food Microbiol. 31 March 2009;130(2):77-87.

7. O'Brien SJ, Gillespie IA, Sivanesan MA, Elson R, Hughes C, Adak GK. Publication bias in foodborne outbreaks of infectious intestinal disease and its implications for evidence-based food policy. England and Wales 1992-2003. Epidemiol Infect. 2006 Aug;134(4):667-74. PubMed PMID:16420723.

8. World Health Organization, ed. Foodborne disease outbreaks: guidelines for investigation and control. World Health Organization, 2008.

9. Regattieri A, Gamberi M, Manzini R. Traceability of food products: General framework and experimental evidence. J Food Engineering. July 2007;81(2):347-356,

10. Schwägele F. Traceability from a European perspective. Meat Science. September 2005;71(1):164-173.

11. Kelly S, Heaton K, Hoogewerff J. Tracing the geographical origin of food: The application of multi-element and multi-isotope analysis. Trends in Food Science & Technology. December 2005;16(12):555-567.

12. Keeling MJ, Eames KT. Networks and epidemic models. J R Soc Interface. 2005 Sep 22;2(4):295-307. PubMed PMID:16849187.

13. Brockmann D, David V, Gallardo AM. Human mobility and spatial disease dynamics. Reviews of Nonlinear Dynamics and Complexity. 2009;2:1-24.

14. Riley S. Large-Scale Spatial-Transmission Models of Infectious Disease. Science. 1 June

2007;316(5829):1298-1301.

15. Hufnagel L, Brockmann D, Geisel T. Forecast and control of epidemics in a globalized world. Proc Natl Acad Sci U S A. 2004 Oct 19;101(42):15124-9. PubMed PMID:15477600.

16. Pérez-Reche FJ, Neri FM, Taraskin SN, Gilligan CA. Prediction of invasion from the early stage of an epidemic. J R Soc Interface. 2012 Sep 7;9(74):2085-96. PubMed PMID:22513723.

17. Shah D, Zaman T. Detecting sources of computer viruses in networks: theory and experiment. In: Proceedings of the ACM SIGMETRICS'10. 201:203-214.

18. Shah D, Zaman T. Rumor centrality: A Universal Source Detector. In: Proceedings of the ACM SIGMETRICS'12. 199-210.

19. Pinto PC, Thiran P, Vetterli M. Locating the Source of Diffusion in Large-Scale Networks. Phys Rev Lett. August 2012;109(6):068702.

20. Lappas T, Terzi E, Gunopulos D, Mannila H. Finding effectors in social networks. In: Proceedings of the ACM SIGKDD'10. 1059-1068.

21. Milling C, Caramanis C, Mannor S, Shakkottai S. On identifying the causative network of an epidemic. In: Proceedings of Communication, Control, and Computing Annual Allerton Conference. October 2012:909-914.

22. Prakash BA, Vreeken J, Faloutsos C. Spotting Culprits in Epidemics: How Many and Which Ones? In: Proceedings of 12th International Conference on Data Mining. 2012:11-20.

23. V. Fioriti and M. Chinnici. Predicting the sources of an outbreak with a spectral technique, 2012.
REFERENCE LINK

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

25. Brockmann D, Helbing D. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. Science. December 2013;342(6164):1337-1342.

26. Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, et al. Epidemic Profile of Shiga-Toxin Producing Escherichia coli O104:H4 Outbreak in Germany. New England Journal of Medicine. 2011;365(19):1771–1780.

27. Buchholz U, Bernard H, Werber D, B¨ hmer MM, Remschmidt C, Wilking H, et al. German Outbreak of Es- o cherichia coli O104:H4 Associated with Sprouts. New England Journal of Medicine. 2011;365(19):1763–1770.

28. Robert Koch-Institute. SurvStat@RKI.de; 2012.
REFERENCE LINK

29. Bundesamt für Kartographie und Geodäsie. GEO84 Verwaltungsgrenzen; 2010.
REFERENCE LINK

30. Anderson J. A Theoretical Foundation for the Gravity Equation. American Economic Review. 1979;69:106–116.

31. Haag G, Weidlich W. A Stochastic Theory of Interregional Migration. Geographical Analysis. 2010 Sep;16(4):331–357.

32. Kaluza P, Kölzsch A, Gastner MT, Blasius B. The complex network of global cargo ship movements. J R Soc Interface. 2010 Jul 6;7(48):1093-103. PubMed PMID:20086053.

33. Min Y, Chang J, Jin X, Zhong Y, Ge Y. The role of vegetables trade network in global epidemics; 2011.
REFERENCE LINK

34. Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature. 2006;439(7075):462–465.

35. Gonzalez MC, Barabasi AL. Understanding individual human mobility patterns. Nature. 2008;453(7196):779-782.

36. Serrano MA, Boguñá M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci U S A. 2009 Apr 21;106(16):6483-8. PubMed PMID:19357301.

37. Woolley-Meza O, Thiemann C, Grady D, Lee JJ, Seebens H, Blasius B, et al. Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. The European Physical Journal B. 2011 Dec;84(4):589–600.
REFERENCE LINK

38. Anderson RM, May RM. Infectious diseases of humans : dynamics and control. Oxford New York: Oxford University Press; 1991.