Accelerating investigation of food-borne disease outbreaks using pro-active geospatial modeling of food supply chains

Daniel Doerr, Kun Hu, Sondra Renly, Stefan Edlund, Matthew Davis, James H Kaufman
IBM Almaden Research Center
650 Harry Road,
San Jose, CA USA
khu@us.ibm.com

Justin Lessler
Department of Epidemiology
Johns Hopkins Bloomberg School of Public Health,
Baltimore, MD, USA
jlessler@jhsph.edu

Matthias Filter, Annemarie Käsbohrer, Bernd Appel
Federal Institute for Risk Assessment
Diedersdorfer Weg 1,
Berlin, Germany
matthias.filter@bfr.bund.de

ABSTRACT
Over the last decades the globalization of trade has significantly altered the topology of food supply chains. Even though food-borne illness has been consistently on the decline, the hazardous impact of contamination events is larger [1-3]. Possible contaminants include pathogenic bacteria, viruses, parasites, toxins or chemicals. Contamination can occur accidentally, e.g. due to improper handling, preparation, or storage, or intentionally as the melamine milk crisis proved. To identify the source of a food-borne disease it is often necessary to reconstruct the food distribution networks spanning different distribution channels or product groups. The time needed to trace back the contamination source ranges from days to weeks and significantly influences the economic and public health impact of a disease outbreak. In this paper we describe a model-based approach designed to speed up the identification of a food-borne disease outbreak source. Further, we exploit the geospatial information of wholesaler-retailer food distribution networks limited to a given food type and apply a gravity model for food distribution from retailer to consumer. We present a likelihood framework that allows determining the likelihood of wholesale source(s) distributing contaminated food based on geo-coded case reports. The developed method is independent of the underlying food distribution kernel and thus particularly applicable to empirical distributions of food acquisition.

Categories and Subject Descriptors
J.3 [**Computer Applications**]: Life and Medical Sciences. Health. The ACM Computing Classification Scheme: http://www.acm.org/class/1998/

General Terms
Design, Experimentation

Keywords
Food-borne disease, food distribution, geospatial data, geospatial modeling, maximum likelihood estimation.

ACM SIGSPATIAL HealthGIS'12, November 6, 2012. Redondo Beach, CA, USA

1. INTRODUCTION

1.1 Global situations
Globalization greatly increases the potential damage associated with food contamination events. The occurrence of various international food safety events over the last few decades undermines public confidence in food safety [1]. Yet, relevance and public interest in food safety issues is endorsed through the establishment of various international food safety standards. This research addresses the question whether there are computational techniques that help to identify possible sources of contamination in the early stages of a disease outbreak, or even make pro-active predictions on likely contamination sources before the onset of an outbreak. Early identification of emerging food-mediated disease outbreaks is a key to prevent them from developing into large-scale health risks. Early identification can reduce the magnitude of new outbreaks and support public health response strategies. Geographic information systems (GIS) can support public health reporting systems to gather information for pro-active investigations. Spatial information of each component in the food distribution and supply chains can be used to define the network relationship between sources of contaminated food, wholesalers, retailers and consumers (and subsequent public health case reports).

1.2 Specific challenge
The need for targeted public health response during an outbreak situation is obvious. Public health investigations are complicated as a result of the large global network of food supply and distribution. In the 2011 European EHEC outbreak, the epidemic had almost subsided before the investigation was completed to identify the source of contamination (*Escherichia coli* O104:H4 in bean sprouts seeds imported from Egypt) [2]. However, many types of food sources are reportedly affected by contamination events. The prioritization of food products that should be closely monitored with respect to specific hazards is a challenge for risk assessors in the public and private sectors. This includes both evaluations of risk factors and the identification of uncertainties on humans, animals, or the environment [3, 4].

1.3 The goals of this paper
This paper describes a method to pro-actively shorten the time required for public health investigation once an outbreak occurs. The method involves creating a probability distribution that describes hypothetical food consumption scenarios within a synthetic population by food types and food distribution sources

(i.e. the wholesaler or distributor). This probability distribution can then be used for maximum likelihood estimation (MLE) to quickly identify the most likely wholesale source of contaminated food once an outbreak occurs. We demonstrate and test the method using a synthetic dataset and discuss the influence of model parameters on its performance.

2. METHOD

2.1 Models of food supply and consumption network

Modeling a food-borne disease outbreak requires at least a number of independent geospatial sub-models and datasets: (1) a model of food distribution by wholesalers, (2) a population model describing the geospatial distribution of the human population, (3) a model describing consumer shopping behavior for food. This latter model defines the distribution of food within the population.

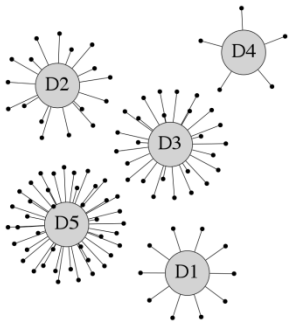


Figure 1: Distributors (D1-5) are associated with retail stores following a power law. In this graphical illustration, retail stores are represented by small black dots at the end of each edge incident to the distributors' nodes.

2.2 Food transportation and retailers network

Wholesalers deliver food to retail stores and restaurants where nearby populations purchase/consume food products. Although a wholesaler is not associated with a specific geo location, retail stores and restaurants are. Considering an international food market, it is reasonable to assume that the transfer sites of wholesalers have negligible influence on the spatial distribution of their associated retail stores. Commonly, the graph connecting wholesalers with retail stores is modeled as scale-free network [5, 6]. In a scale-free network, the node degree, d, follows a power law: $d \sim x^{-\kappa}$ (see example in Figure 1).

In our simple model, we assume that each wholesaler is directly connected to one or more retailers. Again, the number of retail stores associated with a wholesaler follows a power law (see Figure 1). For a specific food type, we assume in our simple model that retailers purchase from a single wholesaler (other food types may come from other wholesalers).

2.3 Population model

The population model drives the food demand and thus determines the geospatial probability distribution for possible infections. In our experiments, we applied a simplified model of population distribution by choosing a uniform distribution of spherical population centers. Each population center is assigned with a random population density. This population density declines linearly with the distance to the center (see example in Figure 2). The analysis described here can easily be applied to other population (demographic) models including models derived from real census data.

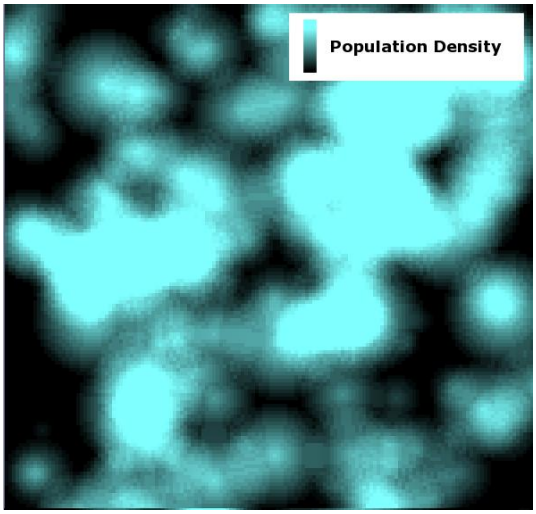


Figure 2: Random sample of a population on a 150×150 grid with 100 population centers. Brighter cyan color represents a higher population density compared to dark areas.

2.4 Consumer grocery shopping behavior

Arguably, access to empirical distributions of food consumption is most providential in investigating food-borne disease outbreaks. However, such data is often not available. Nevertheless, it can be estimated based on various models of shopping behavior. Gravity models have been tested in numerous studies to describe grocery shopping behavior [7-10]. For a given *residence location*, i, Huff's gravity model describes the probability of shopping at retail store j as a function of the retail store's *attraction factor* A_j and the relative distance between i and j, $d_{ij}^{-\gamma}$ [7]. The distance decay rate (the power law exponent) γ reflects the influence of travel distance in the shopping trip distribution:

$$P_{ij} = \frac{A_j \cdot d_{ij}^{-\gamma}}{\sum_k \left(A_k \cdot d_{ik}^{-\gamma}\right)}.$$

Increasing γ reduces the average distance people are likely to travel in purchasing food. The exponent may depend on food type, for example. In the literature studies have derived γ in the range $2 \leq \gamma \leq 3$ [7, 9]. The attraction factor A_j is intended to capture the importance or weight of the store j. Consumer's shopping behavior differs in the types of product, the consumers' income, the retailers' proximity to complementary shops or locations of the customer's other activities such as workplace, gym, etc. This is an effect known as *trip chaining*. Based on Huff's probabilistic formulation, the *multiplicative competitive interaction model* (MCI) [8] provides a sensible generalization that allows for an arbitrary number of parameters other than the attraction factor and the distance decay rate. For the current study, the experimental results presented are confined to the parameters of the basic Huff's model [7].

2.5 Food consumption events

For simplicity, we assume that each individual in the population consumes the same amount of food (homogeneity of the population). Similarly, in each time step, each individual buys food at a single retail store based on the gravity model described above. Since the distribution of food and the population density is known, consumption events can be sampled accordingly. In modeling a food-borne disease outbreak, a single wholesaler is affected and a certain fraction of its conveyed food is contaminated. By sampling a consumption event of contaminated food, we can simulate reported cases of infection. In this study we neglect background gastro-intestinal illness and assume that every reported case is "correct".

2.6 Predicting source of contamination by maximum likelihood estimation

Based on maximum likelihood estimation (MLE), we can determine the likelihood that each distributor is distributing the contaminated food. The MLE can be directly calculated under the presented assumptions of our model.

2.7 MLE for observing a single reported case

Let $\theta = \langle 0, \dots, 0, 1, 0, \dots, 0 \rangle$ be a vector in which the jth entry denotes the likelihood that wholesaler j's food is contaminated. For a single reported case i we aim to estimate the following likelihood:

$$\mathcal{L}(\theta \mid i \text{ is infected and } i \text{ lives at } x_i, y_i)$$

$$= \mathrm{P}(i \text{ is infected and } i \text{ lives at } x_i, y_i \mid \theta)$$

$$= \mathrm{P}(i \text{ lives at } x_i, y_i) \cdot \mathrm{P}(i \text{ is infected} \mid i \text{ lives at } x_i, y_i, \theta)$$

$$= \varphi_{x_i,y_i} \cdot \prod_j [P(i \text{ bought from } j \mid i \text{ lives at } x_i, y_i)]^{\theta_j},$$

where φ_{x_i,y_i} denotes the population density at point x_i, y_i. Note that under the assumption that there is exactly one wholesaler k distributing contaminated food,

$$\mathcal{L}(\theta \mid i \text{ is infected and } i \text{ lives at } x_i, y_i)$$

$$= \varphi_{x_i,y_i} \cdot [P(i \text{ bought from } k \mid i \text{ lives at } x_i, y_i)] \cdot \underbrace{\prod_{j \neq k} [P(i \text{ bought from } j \mid i \text{ lives at } x_i, y_i)]^{\theta_j}}_{=1}.$$

Each wholesaler j is associated with a set of retail stores R_j:

$$\mathcal{L}(\theta \mid i \text{ is infected and } i \text{ lives at } x_i, y_i) = \varphi_{x_i,y_i} \cdot \prod_j \left[\frac{1}{|R_j|} \sum_{k \in R_j} P(i \text{ bought at } k \mid i \text{ lives at } x_i, y_i) \right]^{\theta_j}.$$

2.8 MLE for several reported cases

We now extend the previously developed MLE to a data set D including several reported cases:

$$\mathcal{L}(\theta \mid D) = \prod_{i \in D} \left\{ \varphi_{x_i,y_i} \cdot \prod_j \left[\frac{1}{|R_j|} \cdot \sum_{k \in R_j} P(i \text{ bought at } k \mid i \text{ lives at } x_i, y_i) \right]^{\theta_j} \right\}.$$

Let $f(i,k)$ denote the probability that i shops at store k given i lives at x_i, y_i:

$$O(\theta \mid D) = \sum_{i \in D} \left\{ \log \varphi_{x_i,y_i} + \sum_j \theta_j \log \left[\frac{1}{|R_j|} \cdot \sum_{k \in R_j} f(i,k) \right] \right\}.$$

Since φ_{x_i,y_i} is the same for all distributors, it can be discarded when optimizing for θ:

$$O(\theta \mid D) = \sum_{i \in D} \sum_j \theta_j \log \left[\frac{1}{|R_j|} \cdot \sum_{k \in R_j} f(i,k) \right].$$

3. RESULTS AND DISCUSSION

3.1 Experiment procedure

We conducted several experiments to evaluate the performance of the ML method. The effect of different parameters of our model is evaluated by the success rate of the presented ML method in each experiment setting. All experiments are performed under the same population model using a 100×100 grid with 5 distributors and 100 retail stores. Each data point is averaged over 100 experiments. For each run, the population density is newly generated. Similarly, associations between distributors and retail stores are re-sampled.

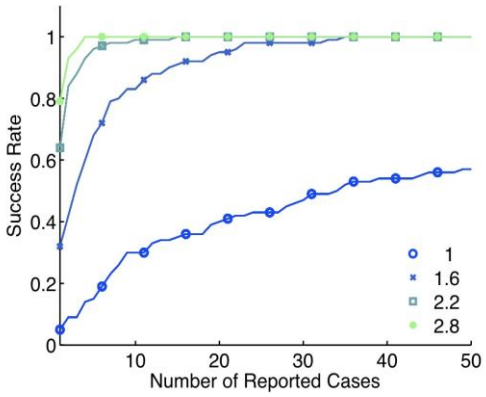


Figure 3: Success rate of ML as a function of the number of reported cases. The plot shows experimental results for various decay rates $\boldsymbol{\gamma} \in \{\mathbf{1}, \mathbf{1.6}, \mathbf{2.2}, \mathbf{2.8}\}$ for a fixed scale-free network exponent of $\boldsymbol{\kappa} = \mathbf{1}$.

3.2 Success rate

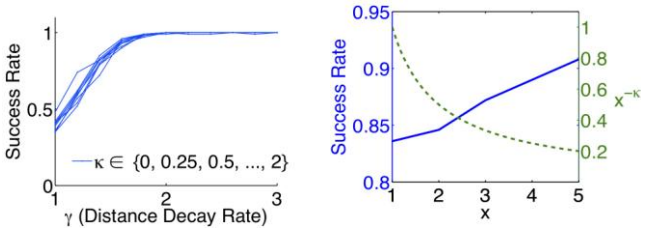
The experimental results depicted in Figure 3 reveal the rate of success of the ML method in determining the true distributor of contaminated food with respect to the number of reported cases. The experiment was repeated for several values of the distance decay rate γ shown in the legend of the figure. The results demonstrate a high success rate for realistic values of $\gamma \geq 2$, even for a low number of reported cases. It is noted that the success rate is almost saturated for the number of reported cases $\gtrsim 30$ and decay rate exponent $\gtrsim 1.6$. The method performs well over the literature range for the distance exponent $2 \leq \gamma \leq 3$ [7, 9].

3.3 Distance decay rate

Empirical studies of consumer's shopping behavior reveal that the distance decay rate depends on the type of consumption goods. For example, consumers are willing to travel longer distances to purchase clothing than to shop for groceries [9]. Consequentially, we hypothesize that the distance decay rate is specific for certain kinds of foods. That is, people are willing to travel longer distances to shop for exotic foods at a specialty shop as opposed to ordinary goods such as dairy products. In the face of the increasing popularity of organic food, this factor may be crucial in fine-tuning food type specific models of food-borne outbreaks. Our simulations indicate that the distance decay rate has a crucial influence in the performance of the ML method (shown in Figure 4).

3.4 Scale-free network dynamics

Recall that the association between wholesalers and retail stores depends on the scale-free network exponent κ (refer to subsection 2.2). Varying this parameter has no apparent influence on the overall success rate of the ML method in our simulations, if the choice of the contaminated wholesaler is uniform among all experimental runs as depicted in Figure 5. However, the performance of ML depends on the node degree of the contaminated distributor. Figure 4 (b) shows that the success rate for small distributors with few retail stores is higher than for large distributors.



(a) **(b)**

Figure 4: **(a)** Success rate of ML as a function of the distance decay rate $\boldsymbol{\gamma}$ for a fixed number of 20 reported cases. The experiment has been repeated for several values of $\boldsymbol{\kappa} \in \{\mathbf{0}, \mathbf{0.25}, \mathbf{0.5}, \ldots, \mathbf{2}\}$. (b) Success rate (blue solid line) of ML as a function of the power law distribution for $\boldsymbol{\kappa} = \mathbf{1}$ (green dotted line) and 20 reported cases. The success rate is averaged over 500 experimental runs with fixed $\boldsymbol{\gamma} = \mathbf{2}$.

4. CONCLUSION

This work proposes a method, given a set of geospatial case reports, to identify the wholesaler(s) most likely involved in the distribution of contaminated food. The proposed likelihood framework is independent of the underlying distribution of food consumption. In this work, we present an integrated model of the food supply chain including the distribution network of wholesalers, the geospatial distribution of population density and retailers' locations, and consumer behavior in choosing retail stores for grocery shopping. The method of maximum likelihood (MLE) is used to predict the most likely contaminated wholesaler. In addition, we conducted numerical experiments using multiple simulations to measure the accuracy of the technique. Therein we study the influence of several parameters of our model in the success rate of the ML method.

4.1 Future work

Our approach can be extended and improved by incorporating empirical distributions of population demographics and food consumption. Modeling food consumption based on a gravity model, important demographic factors (e.g., income) that impact shopping behavior (possibility of travel between workplace and shops, etc) should be included. Consumer behavior should also be differentiated based on food type. Temporal (real) data can be included to further differentiate food distribution by wholesale source. Harvesting periods of farms in different geographic locations, production cycles of manufacturers, and delivery schedules of wholesalers all vary with time. Temporal variation can be used to improve the accuracy in predicting the contaminated wholesaler and food type.

The robustness of the ML approach against model error must be studied. That is, reported cases of infection are reported as confirmed or presumptive. A presumptive report may rarely be confirmed in practice, meaning the "correctness" of infection may never be validated. Further, our hypothetical scenarios of food-borne diseases can be extended to the case where contamination occurs during processing or at the retail store itself, affecting only a sub-graph of the distribution network. The scenario considered in this paper is a simple case (model). More complicated food distribution network, shopping behavior from real data may lead to different conclusions

5. REFERENCES

[1] Marvin, H. J. P., Kleter, G. A., Frewer, L. J., Cope, S., Wentholt, M. T. A. and Rowe, G. A working procedure for identifying emerging food safety issues at an early stage: Implications for European and international risk management practices. Food Control, 20, 4, 2009, 345-356.

[2] Wichmann-Schauer, H., Hiller, P. and Reinecke, A. EHEC-Ausbruch 2011: Ausbruchsaufklärung entlang der Lebensmittelkette. UMID.

[3] WHO Food Safety-Risk Assessment, http://www.who.int /foodsafety/micro/riskassessment/en/. 2012.

[4] EC Directive 2001/18/EC of the European Parliament and of the Council of 12 March 2001 on the deliberate release into the environment of genetically modified organisms and repealing Council Directive 90/220/EEC. Official Journal of the European Communities, L1062001, 1-39.

[5] Barabási, A.-L. and Albert, R. Emergence of Scaling in Random Networks. Science, 286, 5439, October 15, 1999 1999, 509-512.

[6] Ding, Q. and Wang, X. On Complexity Structure of Supply Chain Network, 2009.

[7] Huff, D. L. A probabilistic analysis of shopping center trade areas. Land Economics, 39, 1, 1963, 81-90.

[8] Nakanishi, M. and Cooper, L. G. Parameter Estimation for a Multiplicative Competitive Interaction Model: Least Squares Approach. Journal of Marketing Research, 11, 3, Aug 01 1974, 303-311.

[9] Kubis, A. and Hartmann, M. Analysis of Location of Large-area Shopping Centres. A Probabilistic Gravity Model for the Halle–Leipzig Area. Jahrbuch für Regionalwissenschaft, 27, 1, Feb 19 2007, 43-57.

[10] Veenstra, S. A., Thomas, T. and Tutert, S. I. A. Trip distribution for limited destinations: a case study for grocery shopping trips in the Netherlands. Transportation, 37, 4, 2010, 663-676.