

A Bayesian Network for Outbreak Detection and Prediction

Xia Jiang, Garrick L. Wallstrom

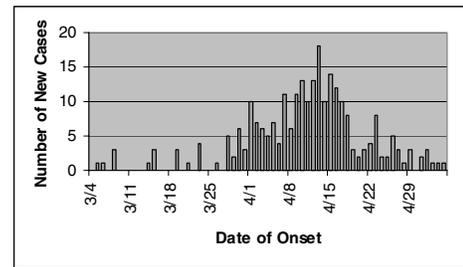
Real-time Outbreak and Disease Surveillance (RODS) Laboratory
University of Pittsburgh, Pittsburgh, PA
xjiang@cbmi.pitt.edu, garrick@cbmi.pitt.edu

Abstract

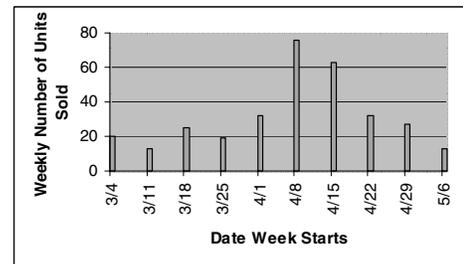
Health care officials are increasingly concerned with knowing early whether an outbreak of a particular disease is unfolding. We often have daily counts of some variable that are indicative of the number of individuals in a given community becoming sick each day with a particular disease. By monitoring these daily counts we can possibly detect an outbreak in an early stage. A number of classical time-series methods have been applied to outbreak detection based on monitoring daily counts of some variables. These classical methods only give us an alert as to whether there may be an outbreak. They do not predict properties of the outbreak such as its size, duration, and how far we are into the outbreak. Knowing the probable values of these variables can help guide us to a cost-effective decision that maximizes expected utility. Bayesian networks have become one of the most prominent architectures for reasoning under uncertainty in artificial intelligence. We present an intelligent system, implemented using a Bayesian network, which not only detects an outbreak, but predicts its size and duration, and estimates how far we are into the outbreak. We show results of investigating the performance of the system using simulated outbreaks based on real outbreak data. These results indicate that the system shows promise of being able to predict properties of an outbreak.

Introduction

(Le Strat and Carrat 1999) define an **epidemic** as ‘the occurrence of a number of cases of a disease, in a given period of time in a given population, that exceed the expected number.’ (Last 2000) defines an **outbreak** as ‘an epidemic limited to localized increase, e.g., in a village, town, or institution.’ Health care departments are increasingly concerned with knowing early whether an outbreak of a particular disease is unfolding. If we can detect an outbreak and estimate its potential cost early, we can take appropriate measures (relative to its potential cost) to control it. An **epidemic curve** (epi curve) is a graphical depiction of the number of outbreak cases by date of onset of illness. Figure 1 (a) shows an epi curve constructed from a sample of the population affected by a *Cryptosporidium* outbreak in North Battleford, Saskatchewan in spring, 2001. In Figure 1 (a) the counts up



(a)



(b)

Figure 1: An epi curve for a *Cryptosporidium* outbreak in North Battleford is in (a), while weekly OTC sales of anti-diarrheal drugs at one pharmacy in North Battleford is in (b). These curves were constructed from data obtained from (Paulson 2001).

until about March 20 are **background counts**, which are the counts when no outbreak is taking place.

The epi curve for an outbreak is often mirrored in the daily counts of some observable variable(s). Consider again the epi curve for the *Cryptosporidium* outbreak shown in Figure 1 (a). *Cryptosporidium* infection causes diarrhea. Figure 1 (b) show the weekly over-the-counter (OTC) sales of anti-diarrheal drugs at one pharmacy in North Battleford. The correlation between these two curves indicates that by monitoring OTC sales of such drugs we can possibly detect a *Cryptosporidium* outbreak in an early stage, and then take some preventative measure such as issuance of a boil water alert.

A number of classical time-series methods have been ap-

plied to outbreak detection. Many are reviewed in (Wong and Moore 2006). A simple method is to take the mean μ and standard deviation σ of the background daily counts, and issue an alert whenever the daily sales exceed μ by $k\sigma$, where k is usually 2 or 3. (Wong and Moore 2006) discuss problems with this simple approach and improvements to it.

The classical methods only give us an alert as to whether there may be an outbreak. They do not predict important properties of the outbreak such as its size and its duration, and they do not estimate how far into the outbreak we are. Knowing the probable values of these variables can help guide us to a decision that maximizes expected utility. We developed a Bayesian network model for an intelligent system that detects outbreaks and predicts their properties. It addresses this problem. Specifically, the network contains nodes/variables for the duration, size, and days since the outbreak began.

In the next section we briefly review Bayesian networks. Then we present the general Bayesian network model for outbreak detection/prediction. After that, we use the model to develop a system that detects *Cryptosporidium* outbreaks, and we show results of experiments using the system.

Bayesian Networks

If we create a DAG (directed acyclic graph) \mathbb{G} whose nodes are random variables, and assume the probability distribution P of the variables in the DAG satisfies the Markov condition with \mathbb{G} , we say we are making the **Markov assumption**, and (\mathbb{G}, P) is called a **Bayesian network** (Neapolitan 2004). A probability distribution P satisfies the **Markov condition** with a DAG \mathbb{G} if the probability of each variable in the DAG is independent of its nondescendants conditional on its parents. There are arguments that a causal DAG satisfies the Markov assumption with the variables in the DAG (See (Pearl 2000).) It is a theorem that if P satisfies the Markov condition with \mathbb{G} , then P is equal to the product of its conditional distributions in \mathbb{G} . So in a Bayesian network we specify the probability distribution by showing the conditional distributions. Efficient inference algorithms have been developed for a large class of Bayesian networks. In particular, inference is usually efficient if the network is sparse. By ‘inference’ we mean we can compute the conditional probability of one or more variables given values of other variables. A **dynamic Bayesian network** is a Bayesian network that explicitly models discrete time. There is a Bayesian network for each time slice, and transition edges and probabilities that relate variable values at a given point in time to values at previous points in time. Efficient inference algorithms for dynamic Bayesian networks exist for a special subclass of networks (See (Neapolitan 2004).).

A Bayesian Network for Outbreak Detection/Prediction

Our intelligent system for outbreak detection represents the problem using a Bayesian network. The general Bayesian network structure for an outbreak detection system appears in Figure 2. We do not show the network as a dynamic Bayesian network for two reasons: 1) We feel it is more

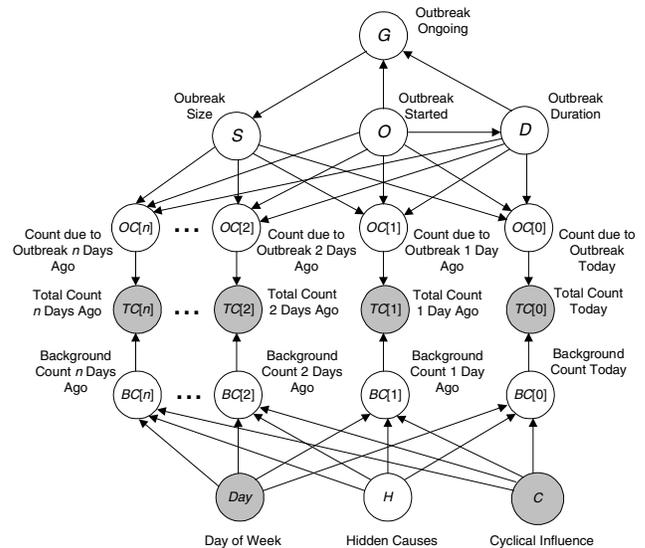


Figure 2: A Bayesian network for outbreak detection/prediction.

lucid to show it as illustrated; 2) In its most general form the dynamic Bayesian network would not be of the type which allows efficient inference. The system would try to predict whether an outbreak of disease X is in progress, along with properties of the outbreak, based on daily counts of variable M . Next we describe each variable in the network.

1. O : This variable represents the number of days ago an outbreak started. It's value is 0 if an outbreak has not started in the previous $durmax$ days, where $durmax$ is the maximum duration of an outbreak of this type.
2. D : The variable D represents the duration of an outbreak that started some time in the previous $durmax$ days. It depends on O because if O is 0, D must be 0.
3. G : The binary variable G represents whether there is an ongoing outbreak today. It depends on O and D because it is 1 if and only if $D \geq O > 0$.
4. S : This is the size of the outbreak, which is the percent of the population that eventually becomes ill due to the outbreak if no measures are taken to control it. It depends on G because if G is 0, S must be 0.
5. $OC[i]$: This is the count of variable M owing to individuals who are sick due to the outbreak. $OC[0]$ is the count today, and $OC[i]$ is the count i days before today. It depends on S and D because larger outbreaks result in larger counts. It depends on O because, during an outbreak, the number of sick individuals increases, as the outbreak progresses, to some peak, and then decreases.
6. Day : This variable represents the day of the week.
7. C : This variable represents some cyclical affect other than day of the week. There could be more than one such variable depending on the application.

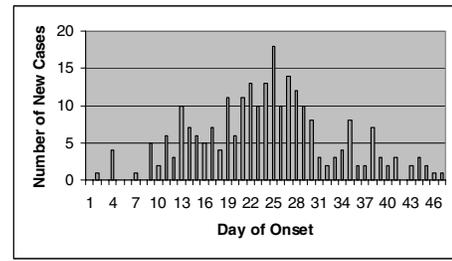
8. H : This variable represents hidden common causes of variable M . It mitigates the relationship between daily counts that are not due to the outbreak. This is not the only way to mitigate this relationship. For example, in a given application it may be best to model it with edges between the variables representing daily counts. (Cheeseman and Stutz 1995) developed AutoClass to learn the range of variables like H from data.
9. $BC[i]$: This is the count of variable M today owing to individuals who are not sick due to the outbreak. This variable depends on H because, for example, it has a higher probability of being large when H is in its high state. It depends on Day and C because we can have different distributions on different days of the week and in different periods of a cycle.
10. $TC[i]$: This is the total count of variable M . It depends on $OC[i]$ and $BC[i]$ because it is the sum of these values.

An Application

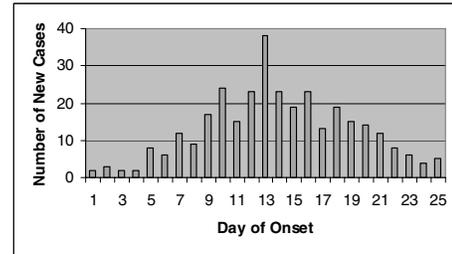
Next we present an application of the model just described to a system that detects/predicts outbreaks of *Cryptosporidium* infection in Cook County, Illinois. *Cryptosporidium* is a water-borne infection that causes diarrhea. The variable observed in the system is OTC sales of antidiarrheal drugs. First we describe the conditional probability distributions in the system.

1. O : We assume that there should be a *Cryptosporidium* outbreak about once every 30 years, and that the duration of such outbreaks is uniformly distributed between 21 and 56 days. Since such outbreaks are so rare, for simplicity we assume there could be at most one in the past 56 days. Therefore, $P(O = 0) = 1 - 56/(30 \times 365)$ and $P(O = i) = 1/(30 \times 365)$ for $1 \leq i \leq 56$.
2. D : If $O = 0$, $P(D = 0) = 1$; otherwise D is uniformly distributed between 21 and 56.
3. G : If $O = 0$ or $D \leq O$, $P(G = 0) = 1$; otherwise $P(G = 1) = 1$.
4. S : If $G = 0$, $P(S = 0) = 1$; otherwise S is uniformly distributed over all integers between 1% and 50%.
5. $OC[i]$: We assume that epi curves for all *Cryptosporidium* outbreaks have the same general shape, and therefore the corresponding curves of OTC sales of antidiarrheal drugs have the same general shape. Figure 3 shows epi curves for the North Battleford, Saskatchewan *Cryptosporidium* outbreak in spring, 2001 and the Milwaukee, Wisconsin *Cryptosporidium* outbreak in spring, 1993. Note that although the outbreaks have different durations, they have the same general shape. Making this assumption, we developed a probability distribution for $OC[0]$ given values of O , D , and S as follows.

We first smoothed the function representing daily OTC sales of antidiarrheal medication during some actual outbreak to obtain a function $g(t)$ whose domain includes all numbers between 1 and the duration D of that outbreak. We chose the North Battleford outbreak mentioned above, and smoothed the sales curve for Pharmacy



(a) North Battleford



(b) Milwaukee

Figure 3: An epi curve for the North Battleford, Saskatchewan *Cryptosporidium* outbreak in spring, 2001 is in (a), while one for the Milwaukee, Wisconsin *Cryptosporidium* outbreak in spring, 1993 is in (b). The curves were respectively constructed from data obtained from (Paulson 2001) and (Mac Kenzie, Hoxie, and Proctor 1994).

A obtained from (Paulson 1994). Next we let D_N and S_N be the duration and size of the North Battleford outbreak, B_N be the average daily background OTC counts in Pharmacy A in North Battleford, and B_C be the average daily background OTC counts in the pharmacies monitored in Cook County. Their values are $D_N = 47$, $S_N = 35.8\%$, $B_N = 1.973$, and $B_C = 986$. Then we set

$$\begin{aligned} \mu(O, D, S) &= \frac{B_C \times D_N \times S}{B_N \times D \times S_N} \times g\left(\frac{(D_N - 1)(O - 1)}{D - 1} + 1\right) \\ &= \frac{986 \times 47 \times S}{1.973 \times D \times 35.8} \times g\left(\frac{46(O - 1)}{D - 1} + 1\right). \quad (1) \end{aligned}$$

For given values of S and D , the result is a function of O whose shape is like $g(t)$ but with duration D , size S , and scaled to Cook County. This formula was developed by (Wallstrom 2006), who offers a formal justification for it. Intuitively, B_C/B_N scales the function to a different sized community, D_N/D scales the function to a different duration, S/S_N scales the function to a different size, and the expression in the argument of g changes the domain of the function from $[1, D_N]$ to $[1, D]$.

Given values of S and D , the function μ is a scaled replica of g . We inserted random fluctuation by letting u be the mean of a negative binomial distribution. We set the dispersion of that distribution equal to 3.52. This value is consistent with the variances we discovered for the background counts.

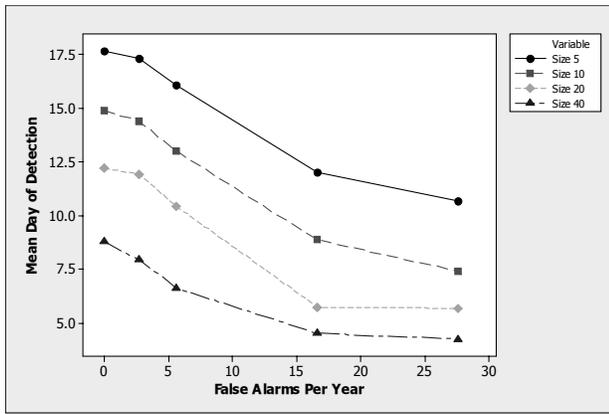


Figure 4: AMOC curves for BayesNet applied to *Cryptosporidium* simulated outbreaks. For each size there were 30 simulations.

6. In this application there is a single variable C that indicates certain holidays. Specifically, the variable was given the value ‘high’ during the 4 days following July 4th and the days from December 26 through January 3.
7. H : Based on manual inspection, we determined three values (low, medium, and high) for background sales data of anti-diarrheal medicine in Cook County.
8. $BC[i]$: For each combination of values of H , Day , and C , we computed a mean and variance of the counts on all days corresponding to those values. We then made the conditional distribution a negative binomial distribution with this mean and variance.

Experiments

We investigated both the detection capability and estimation capability of the system.

Experiment 1

The first experiment investigated how well the system detects outbreaks.

Method It is difficult to test outbreak detection systems because there are so few outbreaks. To address this problem, (Wallstrom, Wagner, and Hogan 2005) developed HIFIDE, which is a system that simulates outbreaks. In particular, the system can simulate OTC sales of anti-diarrheal medication sold during *Cryptosporidium* outbreaks of various sizes. The simulated data are based on data obtained during an outbreak of *Cryptosporidium* in North Battleford. When analyzing outbreak detection algorithms, simulations are ordinarily based only on assumptions about the nature of an outbreak. HIFIDE brings the evaluation one step closer to reality by basing the simulations on a real outbreak.

Using HIFIDE, (Wallstrom, Wagner, and Hogan 2005) evaluated a number of classical outbreak detection systems by injecting simulated *Cryptosporidium* outbreaks into Cook County data between 5/1/2004 and 9/30/2004. So that

Detection Algorithm	Size = 5%	Size = 10%
BayesNet	16.77	13.50
DMSA	19.71	13.55
ARIMA	18.74	14.24
EWMA $w = .01$	16.11	13.44
EWMA $w = .30$	28.72	15.70
CuS-EWMA $w = .01$	14.34	12.63
CuS-EWMA $w = .05$	17.16	14.82
CuS-EWMA $w = .30$	28.62	16.34

Table 1: Mean day of detection with annual false alarm rate equal to 4.

we could compare results of the Bayesian model to the classical methods we simulated outbreaks during that same period. In this way, although the actual simulated data was different, the background counts were the same. The Bayesian network was learned once before the simulations. We simulated outbreaks of duration 47 days (This is the only duration provided by the simulator.) and sizes 5%, 10%, 20%, and 40%. For each size, we did 30 simulations spaced evenly throughout the background period. We used a system that maintained today’s count and the counts on the previous 5 days. We call this system **BayesNet**. The system was developed using Netica (<http://www.norsys.com/>). Inference using Netica took less than 1 second. We issued an alert when the posterior probability of an ongoing outbreak exceeded some threshold. We used AMOC curves to evaluate our results (Fawcett and Provost 1999). In such curves the annual false positive rate is plotted on the x-axis and the mean day of detection on the y-axis.

Results Figure 4 shows AMOC curves for our simulated outbreaks. As would be expected, the detection time decreases as the size of the outbreak increases.

Table 1 compares the results for BayesNet to the results (Wallstrom, Wagner, and Hogan 2005) obtained for various classical methods. The results are only for outbreaks of size 5% and 10% because they did not simulate larger outbreaks. We see that BayesNet compares favorably with these methods. Furthermore, we could have tried different values of the dispersion (Similar to how some classical methods use different weights.) to possibly improve our results, but we did not do this. Figures 5 and 6 compare the results of BayesNet to those of CuSum-EWMA ($w = .05$).

We were able to detect outbreaks earlier using a system that maintained today’s count and the counts on the previous 20 days. We call this system **BayesNet2**. We used BayesNet in the comparisons above so that the AMOC curve technique would compare the system more favorably to the classical methods. BayesNet2 would not compare as favorably due to the following two limitations of AMOC curves: First, suppose that at some threshold a detection system issues an alarm on 3 consecutive days based on certain counts. If we raise the threshold, we might get an alert only on the first day given these same counts. If there is no outbreak, the latter result is preferable because we have fewer false positives. However, if there is an outbreak, the former result is prefer-

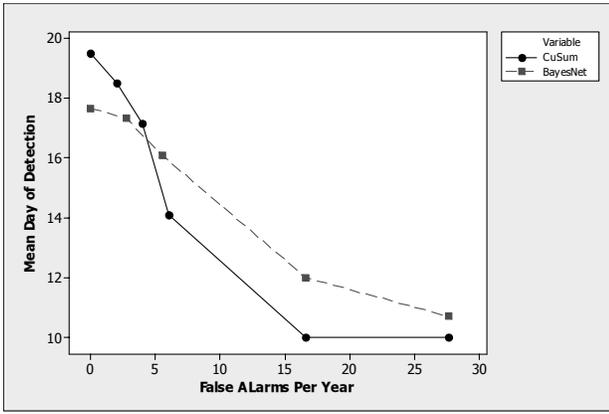


Figure 5: AMOC curves for BayesNet and CuSum-EWMA ($w = .05$) for simulated outbreaks of size 5%.

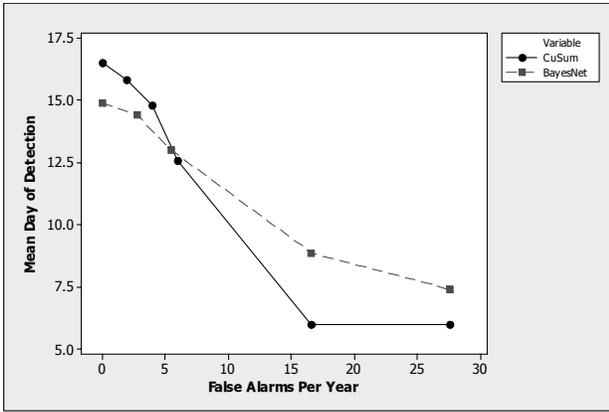


Figure 6: AMOC curves for BayesNet and CuSum-EWMA ($w = .05$) for simulated outbreaks of size 10%.

able because the latter one gives us the false confidence on days two and three that an outbreak is not taking place. So it is not clear which threshold is preferable. AMOC curves only reward systems for the initial day of detection; so they favor the second threshold. The second limitation has to do with systems such as BayesNet which do more than simply issue an alert. Suppose there is no outbreak going on, and the background conditions appear like a small outbreak. On some initial day BayesNet would issue an alert an outbreak is ongoing, on the next day it would still issue an alert but say we are further into the outbreak, and so on. These would all be considered false positives. This problem hurt BayesNet2 in the evaluation because the system would track a false small outbreak throughout much of its duration.

Experiment 2

In this experiment we investigated how well the system estimates and predicts the size (S), duration (D), and days the outbreak has been ongoing (O).

Size	Av S	Er S	Av D	Er D	Av O	Er O
5%	13.55	2.05	43.96	.09	7.04	.56
10%	17.23	.95	43.73	.11	9.43	.83
20%	20.98	.46	44.49	.12	13.75	.81
40%	28.03	.38	45.40	.10	15.34	.65

Table 2: Estimates on the 13th day of the outbreak. The correct value of D is 47 and that of O is 13.

Size	Av S	Er S	Av D	Er D	Av O	Er O
20%	20.26	.40	45.03	.15	20.15	.13
40%	40.00	.13	47.03	.06	20.90	.04

Table 3: Estimates on the 21st day of the outbreak. The correct value of D is 47 and that of O is 21.

Method We used BayesNet2 for this experiment. There were 30 simulations each of sizes 5%, 10%, 20% and 40%. All simulated outbreaks had a duration of 47 days.

Results On a given day of a given simulation, BayesNet2 reports expected values of S , D , and O . Table 2 shows the average (over all 30 simulations) of these values on the 13th day of the simulated outbreaks. The estimates are conditioned on an outbreak taking place. The error (Er) show in that table is the ratio of the square root of the mean square error to the correct value. For example, if the size is 5%

$$Er S = \frac{\sqrt{\sum_{i=1}^{30} (S_i - 5)^2}}{5},$$

where S_i is the expected value of the size reported for the i th simulation. We see that by the 13th day the estimates for size 20% and size 40% outbreaks are fairly good, while those for smaller outbreaks are not. Tables 3 and 4 show that by the 21st day the estimates for size 20% and size 40% outbreaks are quite accurate, as are those for size 5% and size 10% outbreaks by the 25th day.

Figure 7 shows how, in general, the average size estimate improves as we proceed further into the outbreak. Notice that something odd happens in the case of small outbreaks. Namely, at about the time we could actually detect a small outbreak, the average estimate moves in the direction of being too large. This is because a small outbreak on days 10-18 looks much like a larger outbreak on an earlier day.

Discussion

The model presented here is an initial attempt at a system that not only detects outbreaks, but also estimates properties of outbreaks. Our results indicate that it shows promise of being able to estimate these properties.

Size	Av S	Er S	Av D	Er D	Av O	Er O
5%	5.48	.47	45.69	.13	24.30	.11
10%	10.43	.52	44.54	.33	23.41	.16

Table 4: Estimates on the 25th day of the outbreak. The correct value of D is 47 and that of O is 25.

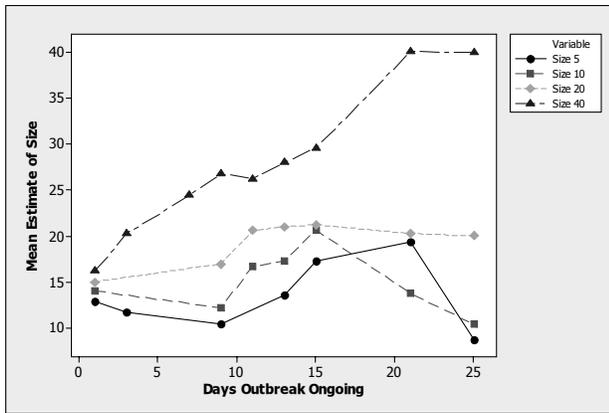


Figure 7: Number of days the outbreak is ongoing is plotted on the x-axis, and the average size estimate is plotted on the y-axis.

Limitations and Future Research

Recall that we obtained the conditional distributions of counts due to the outbreak from data obtained during the North Battleford outbreak, and we evaluated the system using HIFIDE, whose simulations are based on data obtained during that same outbreak. Although such techniques are often used to evaluate detection algorithms, the question remains as to what the performance would be if the model were applied to real data. Recall further that the model assumes epi curves for all *Cryptosporidium* outbreaks have the same shape. Both these matters bear further investigation.

In the case of small outbreaks, the predictions become worse before they become better. It seems that they are sometimes confused with larger outbreaks owing to the fact that outbreak counts during the early days of a small outbreak are quite small. BayesNet2 is memory intensive and the simulations were run on a computer with only 1/2 gigabytes of memory. So it was necessary to maintain counts that were rounded to the nearest 50. The simulations need to be run on a larger computer to see if the results improve. However, it may be that in general it simply is not possible to distinguish small outbreaks on day 13 or so from larger outbreaks at an earlier stage. If this is the case, we still have useful information to provide to the epidemiologist. That is, if a large outbreak is detected, we can state that 'it appears to be a large outbreak in an early stage, but it could be a small outbreak at a later stage. We will know more soon.'

The system described here could readily be modified to handle multivariate times series. Finally, epidemiologists may not be as interested in the potential size and duration of the outbreak as they would be in the potential epi curve. Future research could investigate estimating the epi curve from the estimates of the size and duration.

Related Research

BARD (Hogan et al. 2004) uses a Bayesian network to detect and characterize outbreaks of anthrax due to dispersion of *B. anthracis* spores. It also estimates properties of an out-

break. Specifically, it estimates the geographic scope of an outbreak and the mass of spores released. However, it is designed for a specific type of outbreak, and does not make the predictions discussed here. PANDA (Cooper et al. 2004) is a detection algorithm that models each individual in the population with a Bayesian network. However, it does not predict the properties of an outbreak discussed here.

References

- Cheeseman, P., and Stutz, J. 1995. Bayesian Classification (Autoclass): Theory and Results. In Fayyad, D., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. eds. *Advances in Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press.
- Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W.K., Hogan, W.R., and Wagner, M.M. 2004. Bayesian Biosurveillance of Disease Outbreaks. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Arlington, Virginia: AUAI Press.
- Fawcett, T., and Provost, F. 1999. Activity Monitoring: Noticing Interesting Changes in Behavior. In *Proceedings of the Fifth SIGKDD Conference on Knowledge Discovery and Data Mining*. San Diego, CA: ACM Press.
- Hogan, W., Cooper, G., Wagner, M., and Wallstrom, G. 2004. A Bayesian Anthrax Aerosol Release Detector. Technical Report, RODS Laboratory, Univ. of Pittsburgh.
- Last, J.M. 2000. *A Dictionary of Epidemiology*. New York, NY: Oxford University Press.
- Le Strat, Y., and Carrat, F. 1999. Monitoring Epidemiological Surveillance Data using Hidden Markov Models. *Statistics in Medicine* 18.
- Mac Kenzie, W.R., Hoxie, N.J., and Proctor, M.E. 1994. A Massive Outbreak in Milwaukee of *Cryptosporidium* Infection Transmitted through the Public Water Supply. *The New England Journal of Medicine* 331(22).
- Moore, A., Anderson, B., Kaustav, D., and Wong, W.K. 2006. Combining Multiple Signals for Biosurveillance. In Wagner, M. ed. *Handbook of Biosurveillance*, New York, NY: Elsevier.
- Neapolitan, R.E. 2004. *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.
- Paulson, E., 2001. Waterborne *Cryptosporidiosis* Outbreak, North Battleford, Saskatchewan, spring 2001. *Canadian Communicable Disease Report* 20(1).
- Pearl, J. 2000. *Causality*, New York, NY: Cambridge.
- Wallstrom, G.L. 2006. Probabilistic Interpretation of Surveillance Data. In Wagner, M. ed. *Handbook of Biosurveillance*, New York, NY: Elsevier.
- Wallstrom G.L., Wagner, M.M., and Hogan, W.R. 2005. High-Fidelity Injection Detectability Experiments: A Tool for Evaluation of Syndromic Surveillance Systems. *CDC Morbidity and Mortality Weekly Report*. 54.
- Wong, W.K., and Moore, A. 2006. Classical Time Series Methods for Biosurveillance. In Wagner, M. ed. *Handbook of Biosurveillance*, New York, NY: Elsevier.