A Bayesian Network for Estimating and Predicting Epidemic Curves

Xia Jiang
Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA

OBJECTIVE

This paper describes a Bayesian network model for estimating and predicting an epicurve. Furthermore, it presents results of an experiment testing the accuracy of the predictions.

BACKGROUND

An **outbreak** as 'an epidemic limited to localized increase, e.g., in a village, town, or institution [1].' An **epicurve** is a graphical depiction of the number of outbreak cases by date of onset of illness. The epidemic curve for an outbreak is often mirrored in the daily counts of some observable(s). For example, the epicurve for a flu outbreak is mirrored in the daily over-the-counter (**OTC**) sales of thermometers. The correlation between these curves indicates that by monitoring OTC sales of thermometers we can possibly predict the epicurve for a flu outbreak in an early stage.

Until recently, methods for doing real-time estimation and prediction of the magnitude of an outbreak were very limited. PANDA [2] and BARD [3] provide estimates of some outbreak characteristics such as source and/or route of transmission of the outbreak. A Bayesian network model for outbreak detection and prediction is described in [4]. However, none of these models estimates an epicurve.

METHOD

We developed a Bayesian network model that predicts an epicurve from daily counts of some observable that increase with the size of the outbreak. The network appears in Figure 1. Based on counts of the observable up until today ($C[i]$), the system estimates the epidemic curve values ($E[i]$) on days preceding today, and predicts the epidemic curve values on days following today. The shaded nodes in the networks are constants, while the nodes with shadows are the ones that are instantiated when we use the network to make inference.

RESULTS

Data necessary to estimate flu cases and OTC sales of thermometers were available for Los Angeles, Jefferson County, Orange County, and Allegheny County. Furthermore, these counties had significant flu outbreaks starting in fall, 2003 and lasting 66-68 days. A system was developed using the data from Jefferson County and it was used to predict epicurves

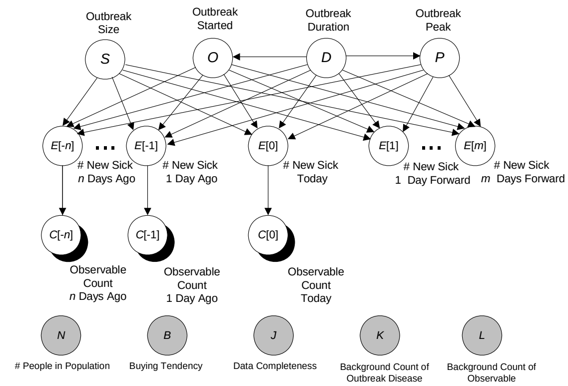

Figure 1. A Bayesian network for predicting epicurves

for all four counties from thermometer sales during the actual outbreaks. The average value of the daily error (relative to a 'gold standard' epicurve) of the estimates obtained on the 20th day, 25th day, and 30th day of the outbreaks are shown in Table 1. We see that in general the predictive accuracy increases as we proceed into the outbreak. However, in the case of Orange County it becomes worse. This is because the tendency to buy thermometers is quite different in Orange County than in Jefferson County. In practice we should actually use Orange County's own buying tendency to develop a system for that County. However, we did not do this in our tests to avoid cyclic testing. Notice that we obtained the best results for Jefferson County, whose buying tendency was used to develop the system.

County	20 Day	25 Day	30 Day
LA	73.7%	39.6%	17.4%
Jefferson	59.6%	26.4%	11.8%
Orange	60.0%	17.0%	35.7%
Allegheny	75.9%	48.7%	36.6%

Table 1. Errors in estimated epicurves

REFERENCES

[1] Cooper, G.F., D. H. Dash, J.D. Levander, W.K. Wong, W.R. Hogan, and M.M. Wagner [2004], "Bayesian Biosurveillance of Disease Outbreaks," *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, July 7-11, 2004, Banff Park Lodge, Banff, Canada.

[2] Hogan, W., Cooper, G., Wagner, M., and Wallstrom, G. 2004. A Bayesian Anthrax Aerosol Release Detector. Technical Report, RODs Laboratory, Univ. of Pittsburgh.

[3] Last, J.M. 2000. *A Dictionary of Epidemiology*, New York, NY: Oxford University Press.

[4] Jiang, X., and G.L. Wallstrom [2006], "A Bayesian Network for Outbreak Detection and Prediction," *Proceeding of AAAI-06*, July, 2006, Boston, MA.