Mining Epidemiological Data Sources in H1N1 Pandemic Using Probabilistic Graphical Models

Masoumeh Izadi David Buckeridge Katia Charland
mtabae@cs.mcgill.ca, david.buckeridge@mcgill.ca, katia.charland@mcgill.ca
Clinical and Health Informatics Research Group
McGill University, 1140 Pine Avenue
Montreal, QC, Canada

***Abstract*—It is generally difficult to estimate disease prevalence or true infection probabilities because these are not observable quantities. However, these parameters can be estimated from available data sources that can provide partial indications of the true incidence of infected cases or prevalence rates. However, building a construct capable of incorporating data from these various sources in a coherent manner is not trivial. In addition, the prevalence of an infectious strain must be estimated in a timely manner. For instance, in an epidemic, this estimate must be obtained within a day or so. We propose to use dynamic Bayesian networks from the class of probabilistic graphical models in order to identify probabilistic relationships between different data streams. This is an initial step towards building a framework that can support data integration and real-time estimation of disease prevalence. Our preliminary results on data sources related to H1N1 pandemic show that the proposed models generalize well.**

Key words– data integration; Bayesian networks; time series analysis; surveillance of infectious disease

I. INTRODUCTION

Infectious disease outbreaks result in high human and financial costs. Respiratory and gastrointestinal infectious diseases, in particular, are among the most prevalent types of infections encountered in routine public health practice. The rapid emergence of the novel pandemic (H1N1) 2009 influenza virus in the spring of 2009 was the most recent example with international concern. This pandemic resulted in more than 18,000 deaths since it appeared in April 2009 [1]. Due to the continued threat of influenza and recognizing the importance of methodological advances to estimate the number of infected cases, building models that provide a good level of understanding of the available data is crucial. Several streams of data such as visits to emergency departments, sales of over the counter drugs, calls to health information lines, and admission to hospitals are routinely used for monitoring outbreaks. In addition, with the advances in research on discovering new sources of data for monitoring of infectious diseases, more emerging data streams become available. However, majority of surveillance systems responsible for monitoring these data treat the sources separately or combine them in an ad hoc fashion.

Combining the data sources can increase statistical power of the data and alleviate biases due to confounding and missing values, in general. Building an architecture to fuse data from different sources in a way that can be easily used for reasoning and prediction is not always easy. Moreover, the desired architecture must be scalable, easily updated, and extensible. Classical approaches to time-series prediction includes linear models such as ARIMA (autoregressive integrated moving average), ARMAX (autoregressive moving average exogenous variables model) [2], [3] and Nonlinear models such as neural networks, decision trees. Problems with these approaches include the fact that it is difficult to incorporate prior knowledge and to integrate multi-dimensional sources into these models. We address this problem using probabilistic graphical models which can be used as appropriate tools for data mining.

Probabilistic graphical models are represented by a graph with nodes and links. The main advantage of these models for data mining and analysis is that the graph structure is used to discover a joint probability distribution for any number of known and unknown quantities simultaneously. Bayesian networks (BNs) and hidden Markov models (HMMs) are among the most popular forms of these models. Both models provide promising methodologies for encoding relations among a large number of random variables based on conditional independence property and are easy to represent real-world problems of high degree of complexity. A generalization over these two models is known as dynamic Bayesian networks (DBNs). DBNs generalize Bayesian networks to model temporal relations and generalize HMMs to model interdependencies between observations.

Our objective is to create a DBN as a unified model to mine different data streams for their interrelationships and to use this model for inference and predictions on data sources used in routine biosurveillance. Another important issue we would like to address is the problem of timeliness. This is specially important in the case of epidemics to have estimates of future counts rapidly. In this paper, we show that there is no need to wait for weeks or even a week in order to estimate the counts of important epidemiological data in future. To further elucidate upon the concept of applicability

 ISBN: 978-1-61208-162-5

of DBNs in this context, a case study is persuaded in this research based on available data sources which carry information related to the infected incidence rate of H1N1 over the pandemic period. For the illustration purposes, in this paper we focus on the data from the island of Montreal, Quebec. Through collaboration with the department of public health in Montreal we had access to data sources such as counts of emergency department visits, calls to health-information lines, vaccination, and hospitalization. There are known qualitative relationships between infection rate or Influenza Like Illnesses (ILI) incidences and a variety of other data sources. For instance, vaccination would reduce the rate of infected cases. Several quantitative relationship between some of these data are also known as domain knowledge. For instance, flu infection makes almost one third of the ILI visits. While very useful, these distributed pieces of information alone are not sufficient to establish a comprehensive model. DBNs are capable of incorporating such domain knowledge in their structure while they build on the knowledge discovered by the data. The steps in the reasoning and prediction by these models will be illustrated through the H1N1 case study in this paper.

II. Dynamic Bayesian Networks

A Bayesian network is a special type of probabilistic graphical models that is represented by a Directed Acyclic Graph (DAG). The DAG explicitly represents independence relationships among random variables. A DAG contains nodes for each random variable and a link between any two statistically correlated nodes. The node originating the directed link is a parent and the terminating node a child. Each node contains a conditional probability table (CPT) that describes the relationship between the node and its parents. If the topology is unknown, i.e., the independence relations among the random variables is unknown, an appropriate structure must be elicited from the data. Automatically learning the structure of a Bayesian network DAG from data is a well-researched but computationally difficult problem [4], [5], [6], [7]. A function is used to score a network with respect to the training data, and a search method is used to look for the network with best score. Different scoring metrics and search methods have been proposed in the literature. The scoring functions used to select models are based on the likelihood function of a model given the data or the logarithm of this function. Since the associated search space is exponentially large, local search-based approaches, which iteratively consider local changes (adding, deleting, and reversing an edge) to the network structure, are usually used to find the best network. This type of search is very useful when dealing with large data sets because of its computational efficiency. One of the most popular search strategies due to its simplicity and good performance [8], [9], [10] in this context is greedy hill-climbing search which starts from an empty graph and gradually improve it by applying the highest scoring single edge addition or removal available. Once the DAG is learned, the parameters of the model (CPTs) need to be specified or directly learned from data. CPTs identify the probabilities of the child being in any specific values given the values of its parents. Parameter learning in Bayesian networks mainly considers maximum likelihood estimation of the model given the data and it is performed through an expectation maximization process. See [7], [11], [6], [12] for parameter learning methods in Bayesian networks. The advantage of DBNs is being able to represent uncertainties, dependencies and dynamics exhibited in different time series. A DBN consists of a finite number of BNs called slices, where each slice corresponds to a particular time instant. BNs corresponding to successive instants are connected through arcs that represent how the state of a random variable changes over time. A DBN is generally assumed to satisfy the Markov property. It is generally assume that the dependencies between the slices of a DBN and their strength do not change over time. Therefore, a DBN can be described by at most a k-slice network (for a k-order Markov domian). DBNs have been applied in a variety of applications from activity recognition and monitoring to medical diagnosis and fault or defect detection. This is the first time that this framework is used for mining in epidemiological data. Ideally, we should be able to learn and discover the probabilistic relationships between data streams through structure learning in DBNs. However, when the system consist of many data streams and in particular when it is partially observed, structure learning in DBNs becomes computationally intensive. This is due to the fact that the space of possible models is so huge that it will be necessary to use strong prior domain knowledge to make the task tractable.

III. Discovering probabilistic relationships between data streams

One practical approach to discover the relationship between time series is to use statistical techniques to learn about the temporal relationships such as lag-lead relationships among the data sources first, combined that with domain knowledge, and then use this information to construct the required DBN. A popular technique in statistics is used for discovering the relationships between time series data or more generally on sequential data, namely: Wavelet Coherency Analysis (WCA) [13], [14].

Wavelet analysis is a useful mathematical technique for analyzing time-series data and periodicities. The wavelet analysis has found many applications in studying longitudinal data [15], [16], [17]. The wavelet coherence is especially useful in highlighting the time and frequency intervals where two time-series have a strong interaction. Such a spectral analysis should be done in an exhaustive way to find the best fit.

 ISBN: 978-1-61208-162-5

The Coherence is defined as the cross-spectrum normalized to an individual power spectrum. It is a number between 0 and 1, and gives a measurement of the cross-correlation between two time-series and a frequency function. The wavelet squared coherency is a measure of the intensity of the covariance of the two series in time-frequency space [18]. It is used to identify frequency bands within which two time series are co-varying. The WCA can provides insight into the temporal relationships to explore in the Bayesian network setting. This is done via the computation of time-frequency maps of the time-variant coherence [15].

IV. Experimental Evaluation

We used and evaluated DBNs in the context of data integration from different sources which partially indicate the pattern of Influenza H1N1 infection. Although, conventionally DBNs are based on first-order Markov processes (i.e. they can be implemented by one-step temporal relationships between two static BNs for only two consecutive time slices), we observed that the data sources we have in hand may potentially indicate more than one step lag between the time series. Therefore, embedding of this particular information into a DBN formulation requires a k-order Markov process for representing a k-layer network, where k indicates the maximum lag between the time-series.

The experiments reported here are based on the data presented in the next section. We learned DBN models from the data in a variety of settings, and compared them with respect to their performance in predicting observable data streams. The main purpose of this phase of the research is to understand how well DBNs can represent the whole processes, how many observations are required, and what sorts of observations are most useful. In all our experiments, we enforced the presence of the arcs in the DBN network structure based on the suggested settings by WCA, or BN structure learning. In performing the BN structure learning, we followed a similar strategy to what suggested by [19]. For each data source, we selected the variables observed at $t, t+1..., t+10$ days and performed hill climbing search to find the network with the best score.

A. Data

Through collaborations with the department of public health in Montreal, we had access to five different data sources. These data sources include: daily counts of emergency department visits (ED), daily counts of calls to health information lines,Info-Sante, (IS), weekly counts of H1N1 vaccination, weekly counts of confirmed cases of H1N1 through lab tests, and weekly counts of admission to the hospitals. Since the data sources have different resolutions in time and have different significance in predicting the number of infected cases, we are only considering the daily time-series of ED and IS, preliminary. Emergency department visits may well estimate incidence of influenza.

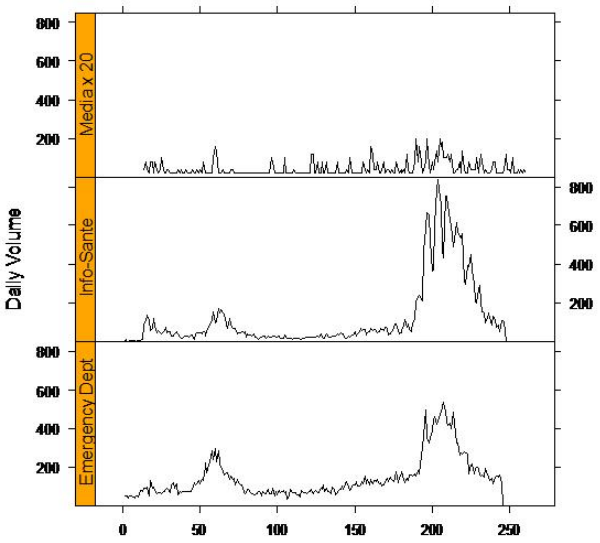

Figure 1. Three data sources of media reports, calls to InfoSante, and emergency department visits from top to bottom.

We can combine emergency department triage data with the telephone survey to characterize the effectiveness of incidence estimation. We aggregated visits for ILI by age group, sex, and day of visit. Similar to the ED data, the IS data can be used to estimate influenza incidence. We aggregated ILI calls by age group, sex, and day of call.

Media reports of deaths from pH1N1 were considered important because of their pronounced effect on the utilization of health services, thus media reports were filtered for content. We also extracted the Media data from the Healthmap [20] on a daily basis.Figure 1 shows the total daily counts of H1N1 media reports about Montreal during the period of April 28, 2009 to December 16, 2009 in the top graph. The second graph illustrate the total daily calls to Info-Sante, and the third graph shows the total daily counts of emergency department visits during the same period. The arrow points to the time when a 13 year old boy (hockey player) in Ontario died on October 26. There were reports of his funeral at around November 4 (t=203 on the time axis). This precedes, by 1 day, the sharp spike in Info-sante calls.

B. Results

The extent of the temporal relationship between IS and ED series data was estimated using WCA in Figure 2. Our results in Figure 2 shows about 2-4 day lead or lag. There is a phase change at around Nov 2, in the second wave. We are able to see a predictable relationship during seasonal influenza (with IS leading ED by approximately 4 days), but during the pandemic (and especially the second wave)

 ISBN: 978-1-61208-162-5

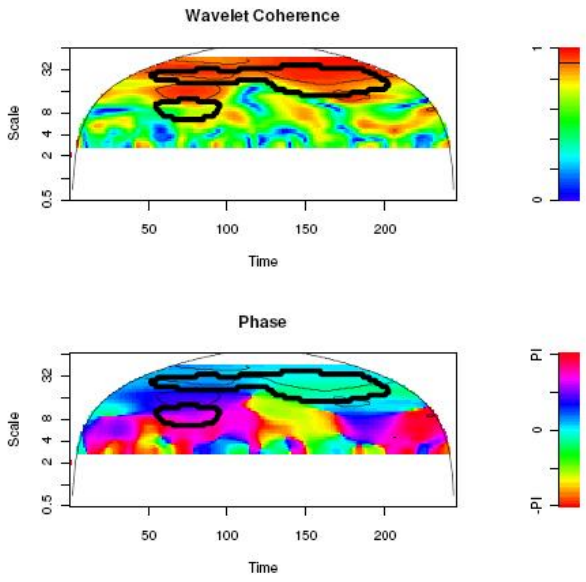

Figure 2. Wavelet coherency analysis for two data sources ED and InfoSante.

the relationship was less predictable. We speculated that it is possibly due to media influence.

In this research, we aimed to learn DBN models that generalize well. The generalization ability of a model G is interpreted as the expected predictive accuracy for the next time series, D_{T+1}. We evaluated the DBN model for prediction accuracy of important observations in time series IS and ED through cross validation techniques. The first set of experiments involved learning DBNs of different complexities. Once trained, we can use the model to do real-time prediction through approximate inference in BNs.

We used a BN structure learning search over the space of all possible graphs to find the best graph, and we discovered two day lag for ED during the seasonal and pandemic flu 2009. However, for an extended period of time (May 1, 2008 to December 30, 2009), which includes non-pandemic, seasonal, and pandemic flu, we found different dependency relations between the two series by BNs structure learning. As the WCA suggested candidate models with 4 days lag, we also tried to train a DBN model with no-phase difference between IS and ED in a DBN (Figure 3). The structure learning method also found that media reports data can lead the Info-Sante data by one day. However, this relationship only exist during the pandemic period in our data sets (April-December 2009). We also presented the Bayesian network models to the experts in public health surveillance and asked them to assess the face validity of the dependence between the time series. The expert feedback was more in favor of IS leading ED.

We experimented with four DBNs that correspond to the settings suggested by BN structure learning and WCA:

- ED leads InfoSante by 2 days
- No phase difference between ED and InfoSante
- InfoSante leads ED by 2 days
- Media leads InfoSante by one day and Infosante leads ED by 2days

Figure 3 shows the unrolled DBNs for seven time steps (weekly). We can treat the unrolled version of a network as a static BN and apply inference algorithms in BNs. We used cross validation for evaluation of all models. Four fifth of the data was used for training and One fifth of the data was used for testing. In each model we provided the information for today's count on ED and IS and predicted the first to 6th next day's counts on both ED and IS. The second model works actually the best when it is trained and tested on the pandemic period (no more than 11% error in predicting ED).

It should be noted that for all models we considered categorization for all variables. This includes Media $\in \{0, 1-3, 3-7, >7\}$, ED $\in \{0-100, 100-200, 200-300, 300-400, 400-500, >500\}$, and IS $\in \{0-100, 100-200, 200-300, 300-400, 400-500, 500-600, 600-700, >700\}$. The results may vary with changing the categorization.

Although, there exist dependencies between the media data and the IS data, we did not see a significant changes in the prediction results for IS. This can be potentially related to other factors which have not been considered in our model or solely related to the experimental setup we selected for these evaluations including the discretization levels of the Media and IS variables and the information provided for reasoning at each time.

V. Conclusions and Future Work

Monitoring epidemiological data is critical for detecting epidemics and for guiding control measures. During the H1N1 pandemic, the Direction de sante publique de Montreal collected data from multiple sources to describe H1N1 influenza infection and associated health care utilization. None of these data sources alone are believed to measure the incidence of H1N1 influenza accurately. In this paper, we proposed a probabilistic graphical model to different heterogeneous data and discover meaningful information these data exhibit. We showed how a DBN model can be used for generating short-term predictions of real-time surveillance data. The estimates are also timely. We showed that only the order of one to two days required in order to estimate future counts in the studied data sources These estimates will be eventually useful in forecasting the spread of H1N1 influenza.

We will continue our investigations for choosing a better DBN structure. We plan to evaluate all lags (plus/minus) 4 and pick the one with the best prediction power. We did not consider the complete DBN model to predict the number of infected cases of H1N1 in this paper. After reaching a good DBN model for integration of data sources, we plan to

 ISBN: 978-1-61208-162-5

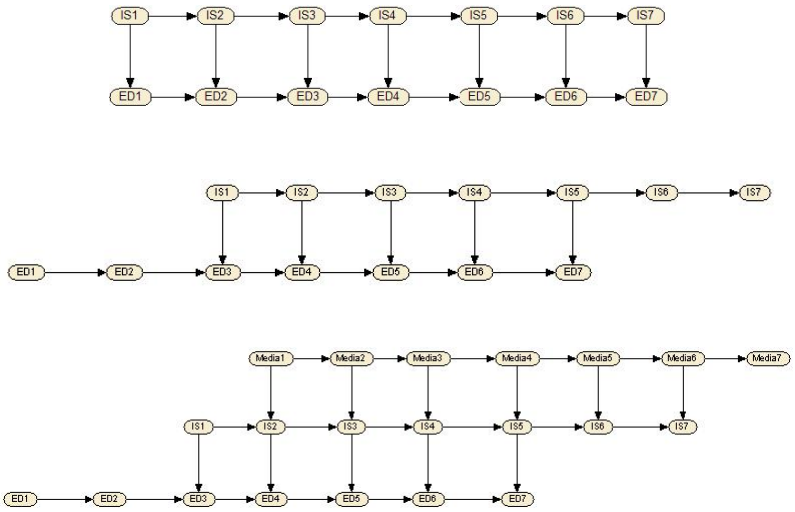

Figure 3. (a) No phase difference between ED and InfoSante, (b) Infosante leads ED by 2days, and (c) Media leads InfoSante by one day and Infosante leads ED by 2days

Table I
COMPARISON OF THE PERFORMANCE OF DIFFERENT DBN MODELS IN PREDICTING INFORMATION-SANTE DATA IN THE NEXT SIX FOLLOWING DAYS.

Model	**error%**					
	Day1	Day2	Day3	Day4	Day5	Day6
ED leads 1-day	19.49	21.19	22.03	26.72	29.03	29.9
IS leads 2-days	18.68	23.37	25.64	26.22	28.81	29.06
Zero-phase difference	18.68	23.37	25.64	26.22	28.81	29.06
Media-effect	18.24	24.21	26.72	27.65	29.31	32.59

Table II
COMPARISON OF THE PERFORMANCE OF DIFFERENT DBN MODELS IN PREDICTING ED DATA IN THE NEXT SIX FOLLOWING DAYS.

Model	**error%**					
	Day1	Day2	Day3	Day4	Day5	Day6
ED leads 1-day	8.47	11.86	14.53	16.1	21.37	24.14
IS leads 2-days	8.47	11.02	12.52	13.33	13.56	18.49
Zero-phase difference	9.32	11.68	12.71	13.64	13.68	16.38

extend the DBN model of observable data sources presented here to what is called an autoregressive hidden Markov models (AHMM) to contain the unobservable infected counts. We can then apply learning algorithms such as Viterbi and Baum-Welch on this hierarchical dynamic Bayesian network just as we can on HMMs to estimate the prevalence of H1N1.

ACKNOWLEDGMENT

The authors acknowledge the contribution of the members of the department of public health in Montreal who agreed to provide data for this research and participated in regular expert meetings. In particular, we would like to thank Lucie Bedard and Robert Allard for invaluable insights for the analysis of our results and in preparation of this paper.

REFERENCES

[1] CIDRAP, "WHO says H1N1 pandemic is over."

[2] H. Burkom, S. Murphy, J. Coberly, and K. Hurt-Mullen, "Public health monitoring tools for multiple data streams," *MMWR*, no. 54, pp. 55–62, August 2005.

[3] B. Reis and K. Mandl, "Time series modeling for syndromic surveillance." *BMC Medical Information Decision Making*, vol. 3, no. 2, 2003.

[4] D. Chickering, D. Heckerman, and C. Meek, "Large-sample learning of bayesian networks is np-hard," *JMLR*, vol. 5, pp. 1287–1330, 2004.

[5] D. Chickering and C. Meek, "Finding optimal bayesian networks," Microsoft Research, Tech. Rep., 2002.

 ISBN: 978-1-61208-162-5

[6] F. A. Jensen, *An Introduction to Bayesian Networks*. Springer, 1996.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[8] I. Tsamardinos, L. Brown, and C. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006.

[9] D. Heckerman, D. Geiger, and D. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, 1995.

[10] N. Friedman, K. Murphy, and S. Russell, "Learning the structure of dynamic probabilistic networks," *Uncertainty in Artificial Intelligence*, pp. 139–147, 1998.

[11] D. Grossman and P. Domingos, "Learning bayesian network classifiers by maximizing conditional likelihood," in *ICML*, 2004.

[12] R. Neapolitan, *Learning Bayesian Networks*. Prentice Hall, 2003.

[13] J. Morlet, G. Arens, I. Foourgeau, and D. Giard, "Wave propagation and sampling theory," *Geophysics*, vol. 47, pp. 203–236, 1982.

[14] S. Mallat, *A Wavelet Tour of Signal Processing*. New York Academic, 1999.

[15] K. Keissar, R. Davrath, and S. Akselrod, "Time and frequency wavelet transform coherence of cardio-respiratory signals during exercise," *Computers in Cardiology*, pp. 733–736, 2006.

[16] T. Li and W. Klemm, "Detection of cognitive binding during ambiguous figure tasks by wavelet coherence analysis of eeg signals," *Pattern Recognition*, pp. 3098–3101, 2000.

[17] A. Grinsted, J. Moore, and S. Jevrejeva, "Application of the cross wavelet transform and wavelet coherence to geophysical time series," *Nonlinear Processes in Geophysics*, vol. 11, pp. 561–566, 2004.

[18] C. Torrence and G. Compo, "A practical guide to wavelet analysis," *Program in Atmospheric and Oceanic Sciences, University of Colorado, Boulder, Colorado*, 1998.

[19] P. Sebastiani, K. Mandl, P. Szolovits, I. Kohane, and M. Romain, "Bayesian dynamic model for influenza surveillance," *Journal of the American Statistical Association*, 2006.

[20] J. Brownstein, C. Freifeld, B. Reis, and K. Mandl, "Surveillance sans frontires: Internet-based emerging infectious disease intelligence and the healthmap project," *PLoS Med*, vol. 5, no. 7, p. 151, 2008.

 ISBN: 978-1-61208-162-5