1

Bayesian Network Scan Statistics for Multivariate Pattern Detection

Daniel B. Neill[1,2], Gregory F. Cooper[3], Kaustav Das[2], Xia Jiang[3], and Jeff Schneider[2]

[1] *Carnegie Mellon University, Heinz School of Public Policy and Management*
[2] *Carnegie Mellon University, School of Computer Science*
[3] *University of Pittsburgh, Department of Biomedical Informatics*

Abstract: We review three recently proposed scan statistic methods for multivariate pattern detection. Each method models the relationship between multiple observed and hidden variables using a Bayesian network structure, drawing inferences about the underlying pattern type and the affected subset of the data. We first discuss the multivariate Bayesian scan statistic (MBSS) proposed by Neill and Cooper (2008). MBSS is a stream-based event surveillance framework that detects and characterizes events given the aggregate counts for multiple data streams. Next, we describe the agent-based Bayesian scan statistic (ABSS) proposed by Jiang and Cooper (2008). ABSS performs event detection and characterization given individual-level data for each agent in a population. Finally, we review the Anomalous Group Detection (AGD) method proposed by Das, Schneider, and Neill (2008). AGD is a general pattern detection approach which learns a Bayesian network structure from data and detects anomalous groups of records.

Keywords and phrases: Pattern detection, event detection, scan statistics, Bayesian networks, biosurveillance

1.1 Introduction

In this chapter, we focus on the problem of *multivariate event surveillance*, in which we monitor multiple data sources with the goal of identifying patterns that correspond to emerging events. More generally, our goal is *pattern detection*: we wish to find subsets of a large, complex dataset that are relevant, either because the group of data records corresponds to some known statistical pattern which we are interested in detecting, or because it is highly anomalous given our current understanding of the data. Here we review three recently pro-

posed Bayesian variants of the spatial scan statistic [Kulldorff (1997)], which extend the scan statistic methodology to enable rapid detection and accurate characterization of events in multivariate datasets. The three methods include the multivariate Bayesian scan statistic (MBSS) method proposed by Neill and Cooper (2008), the agent-based Bayesian scan statistic (ABSS) method proposed by Jiang and Cooper (2008), and the Anomalous Group Detection (AGD) method proposed by Das, Schneider, and Neill (2008). MBSS is a stream-based event surveillance framework that detects and characterizes events given the aggregate counts for multiple data streams, while ABSS performs event detection and characterization given individual-level data for each agent in a population. Finally, AGD is a general pattern detection approach which detects anomalous groups of records in categorical datasets. These methods use Bayesian networks to model the relationship between multiple observed variables, extending the univariate Bayesian spatial scan statistic methodology of Neill *et al.* (2006) to integrate multiple data streams and differentiate between multiple types of event. MBSS and ABSS assume fixed Bayesian network structures, focusing on stream-based and agent-based event surveillance scenarios respectively, while AGD learns the Bayesian network structure from data and can be applied to pattern detection in general multivariate datasets.

1.1.1 Event surveillance

Event surveillance systems monitor massive quantities of multivariate data in order to detect and identify emerging patterns. For example, government agencies responsible for public safety must respond rapidly to potential threats including wars, disease outbreaks, crime waves, natural disasters, and terrorist attacks. Timely and informed responses to such events may substantially reduce the resulting costs to society, while delayed or incorrect responses can have catastrophic results. As a concrete example, we consider the task of disease surveillance, in which we monitor electronically available public health data such as hospital visits and medication sales in order to detect emerging outbreaks of disease. Major health threats such as emerging infectious diseases or bioterrorist attacks require rapid and appropriate responses in order to control the spread of disease and treat infected individuals. However, taking appropriate actions often requires knowledge of the characteristics of the disease (e.g. source, method of transmission, and available treatments) and which areas have been affected. Similarly, serious outbreaks requiring urgent responses must be distinguished from less serious outbreaks (e.g. seasonal influenza) and from irrelevant patterns in the data (e.g. increases in medication sales due to store promotions).

The main goals of event surveillance are to achieve *early detection* and *accurate characterization* of events, identifying which events have occurred and which subsets of the data have been affected by each event. However, the

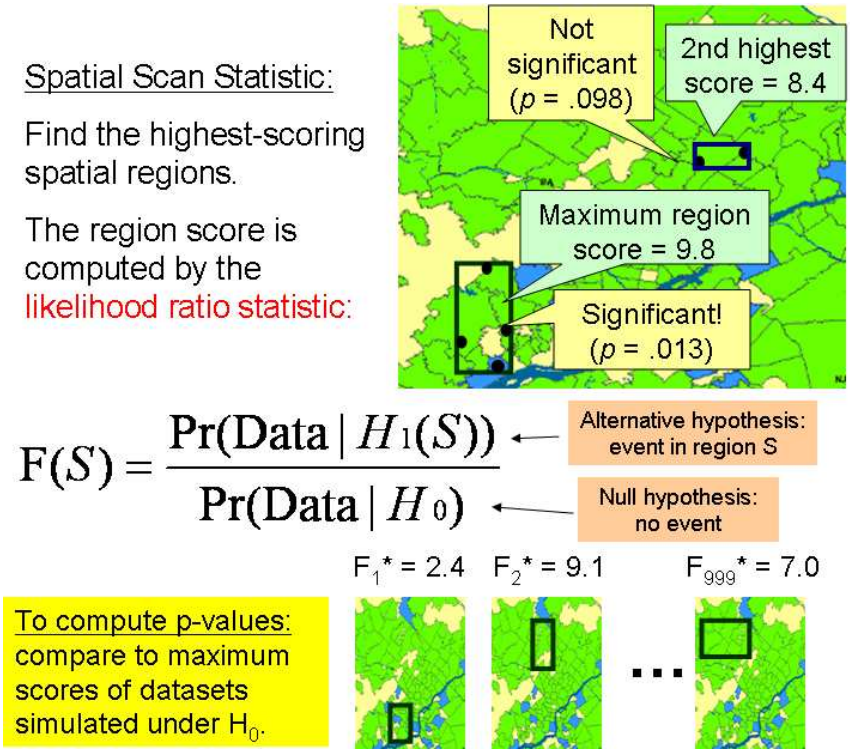

Figure 1.1: Demonstration of the spatial scan statistic.

massive size, high dimensionality, and complex spatial and temporal structure of the multivariate data make these goals difficult to achieve. As discussed by Neill and Cooper (2008), an event surveillance system must meet three general criteria to achieve timely and accurate detection:

1. To achieve high detection power, the system must integrate spatial and temporal information from multiple data streams (or from multiple individuals in a population) in a coherent probabilistic framework, incorporating both prior knowledge and historical data into its models.

2. To achieve accurate characterization of events, the system must be able to model and differentiate between multiple types of events.

3. To achieve a rapid response to emerging events, the system must be computationally efficient, detecting patterns in large real-world datasets in near real time.

We now discuss a variety of commonly used methods for event detection, and consider how well the methods fit these criteria.

1.1.2 The spatial scan statistic

The spatial scan statistic [Kulldorff and Nagarwalla (1995), Kulldorff (1997)] is a well-known method for spatial cluster detection. It is in wide use for monitoring health data, detecting clusters of disease cases due to chronic environmental exposures [Kulldorff *et al.* (1997), Hjalmars *et al.* (1996)], infectious disease outbreaks [Mostashari *et al.* (2003)], or bioterrorist attacks [Neill (2006)]. Given a set of spatial locations s_i, each with a count (e.g. number of disease cases) c_i and an underlying population p_i, the spatial scan finds the most significant clusters by searching over a given set of spatial regions, finding those regions which maximize a likelihood ratio statistic, and computing the statistical significance of the detected regions by randomization testing (Figure 1.1). Assuming that the counts in region S are distributed with some unknown rate of incidence q, the goal of the scan statistic is to find regions where the incidence rate is higher than expected. We can either compare the counts inside and outside region S [Kulldorff (1997)], or alternatively, compare the counts inside region S to their expected values obtained from historical data [Neill *et al.* (2005b)]. In either case, we define the null hypothesis H_0, which assumes no clusters, and the alternative hypothesis $H_1(S)$, which assumes a cluster in region S. We then find the region that maximizes the *likelihood ratio statistic*:

$$F(S) = \frac{\Pr(\text{Data} \mid H_1(S))}{\Pr(\text{Data} \mid H_0)} \tag{1.1}$$

The original presentation of the spatial scan statistic [Kulldorff (1997)] considers two different models, the Bernoulli model and the Poisson model. In the Bernoulli model, each individual is characterized by some binary variable (e.g. whether the individual goes to the Emergency Department with a fever). Under the null hypothesis of no clusters, H_0, every individual has a constant probability q_{all} of having this property, while under the alternative hypothesis of a cluster in region S, $H_1(S)$, the incidence rate is higher inside region S than outside (i.e. $q_{in} > q_{out}$). In the Poisson model, we measure the total count of some event type (for example, the number of over-the-counter cough/cold drugs sold) in each spatial region. Assuming that counts are Poisson distributed with mean proportional to the product of the population p_i and the incidence rate q, we can again compare the rates inside and outside region S. Likelihood ratio statistics for each model are derived by Kulldorff (1997).

While Kulldorff's original spatial scan statistic did not take the time dimension into account, later work generalized this method to the "space-time scan statistic" by considering a time series of counts c_i^t for each spatial location s_i and scanning over variable size temporal windows [Kulldorff *et al.* (1998), Kulldorff (2001)]. Recent extensions such as the expectation-based scan statistic [Neill *et al.* (2005b)] and model-based scan statistic [Kleinman *et al.* (2005)]

also take the time dimension into account by using historical data to model the expected distribution of counts in each spatial location.

Many variants of the spatial and space-time scan statistics have been proposed, differing in both the set of regions to be searched and the underlying statistical models. While Kulldorff's original method [Kulldorff (1997)] assumed circular search regions, other methods have searched over rectangles [Neill *et al.* (2005a)], ellipses [Kulldorff *et al.* (2006)], and various sets of irregularly shaped regions [Duczmal and Assuncao (2004), Patil and Taillie (2004), Tango and Takahashi (2005)]. Similarly, many different statistical models have been considered, ranging from simple Poisson and Gaussian statistics [Neill *et al.* (2005b), Neill (2006)] to robust and nonparametric models [Neill and Sabhnani (2007), Neill and Lingwall (2007)].

Kulldorff *et al.* (2007) recently proposed a multivariate variant of the Poisson spatial scan statistic. This work directly extends the original spatial scan to multiple data streams by assuming that all data streams are independent, thus calculating the likelihood ratio score for a given region as the product of the likelihood ratios for each individual data stream. However, we expect streams to be correlated by spatial and temporal trends and other covariates under the null hypothesis, and by the parameters of an event (e.g. outbreak severity) under the alternative hypothesis. Additionally, Kulldorff's method does not characterize events, differentiate between multiple event types, or incorporate prior information. Nevertheless, it can integrate information from multiple data streams for faster and more accurate detection, and performs well as a "general detector" of anomalous patterns when no prior knowledge of events is assumed. Neill and Cooper (2008) use this method as a baseline for comparison when evaluating the detection power of their MBSS method.

1.1.3 The univariate Bayesian spatial scan statistic

The spatial scan approaches described in Section 1.1.2 fulfill some, but not all, of the criteria for event surveillance discussed above. Spatial scan methods integrate information from multiple spatial locations and multiple time steps, but with the exception of the multivariate Poisson spatial scan [Kulldorff *et al.* (2007)], they can monitor only a single data stream. These methods are also computationally expensive because randomization testing is used to determine the statistical significance of detected clusters, requiring a search over all spatial regions S for many randomly generated datasets. Most importantly, none of these methods can model and differentiate between multiple event types, limiting their usefulness for event characterization.

The Bayesian spatial scan statistic (BSS) method, developed by Neill *et al.* (2006), enables the incorporation of prior information into the event detection process. In the BSS framework, we are given a dataset D, consisting of a time series of counts c_i^t for each spatial location s_i, and we consider a given set of

space-time regions S with prior probabilities $\Pr(H_1(S))$. For some recent past period of time (e.g. the current day), BSS computes the posterior probability that an event has occurred in each spatial region using Bayes' Theorem:

$$\Pr(H_1(S) \mid D) = \frac{\Pr(D \mid H_1(S))\Pr(H_1(S))}{\Pr(D)} \tag{1.2}$$

$$\Pr(H_0 \mid D) = \frac{\Pr(D \mid H_0)\Pr(H_0)}{\Pr(D)} \tag{1.3}$$

The likelihood of the data under each hypothesis is computed using a Gamma-Poisson model, and we can specify a probability distribution for the effects of an event on the affected region S. Neill *et al.* (2006) demonstrated that the Bayesian approach has several advantages over frequentist methods. Computation is much faster in the Bayesian framework since randomization testing is unnecessary, and the results of the BSS method (the posterior probability that each region has been affected) are easy to interpret and visualize. Most importantly, the BSS framework allows us to model the spatial and temporal distribution of events by specifying the region priors $\Pr(H_1(S))$, as well as modeling the effects of an event $H_1(S)$ on the monitored data stream in the affected region S. While the original BSS method only considers a single data stream and a single event type, the recently proposed multivariate Bayesian scan statistic [Neill and Cooper (2008)] extends this framework to multiple streams and multiple types of events. We discuss the MBSS method in more detail in Section 1.2. More generally, the Bayesian framework can be extended to multivariate data by specification of a Bayesian network relating the observed variables and the underlying event. Each of the three methods discussed in this chapter considers a different set of observations and thus assumes a different Bayesian network structure. In the following section, we briefly review Bayesian networks and their application to pattern detection.

1.1.4 Bayesian networks

A Bayesian network [Pearl (1988), Heckerman *et al.* (1995)], or Bayes Net, is a commonly used graphical representation of the joint probability distribution of a set of variables. Bayes Nets are a valuable statistical tool for efficient inference and learning of multivariate probability distributions, and provide a concise and interpretable visualization of the conditional dependencies between variables. They have been used in many anomaly detection applications, including network intrusion detection [Bronstein *et al.* (2001), Ye and Xu (2000)], detecting malicious emails [Dong-Her *et al.* (2004)] and outbreak detection [Wong *et al.* (2003a), 2003b]. Formally, a Bayesian network can be represented as a directed acyclic graph, where each vertex X_i represents a variable, and each edge from a "parent" vertex X_p to a "child" vertex X_c represents the dependence

of X_c on X_p. The joint probability distribution can be concisely expressed as the product of each variable's conditional distribution given the values of that variable's parents: $\Pr(X_1 \ldots X_M) = \prod_{i=1\ldots M} \Pr(X_i \,|\, \text{Parents}(X_i))$. Conditional independencies between variables can also be easily inferred from the network structure: for example, any variable is conditionally independent of its non-descendents given its parents. Inference and learning in Bayesian networks are described in detail by Pearl (1988), Heckerman *et al.* (1995), and many others.

One general approach to anomaly detection using Bayesian networks is to report any individual records with unusually low likelihoods as potential anomalies. In this case, a Bayesian network is learned automatically from a large "training dataset". Established machine learning methods such as Optimal Reinsertion [Moore and Wong (2003)] can be used to efficiently learn the network structure, and the parameters can be optimized by maximum likelihood. We then compute the likelihood of each record in a separate "test dataset" given the Bayes Net model, and report the least likely records. Unlike the scan statistic methods considered here, this method treats each individual data record separately, and does not incorporate any spatial or temporal data or other information about group structure. Das *et al.* (2008) use this method as a baseline for comparison in their evaluation of AGD, as discussed below.

Also relevant to our discussion is the PANDA system for disease surveillance proposed by Cooper *et al.* (2004, 2007), which uses Bayesian network models to differentiate between multiple outbreak types (e.g. the CDC Category A diseases), assuming an underlying agent-based model of Emergency Department visits. Unlike the event detection methods considered here, the baseline version of PANDA-CDCA [Cooper *et al.* (2007)] does not incorporate spatial information, and thus cannot determine which subset of the data has been affected by an event. However, Section 1.3 describes the agent-based Bayesian scan statistic [Jiang and Cooper (2008)], which extends the PANDA-CDCA model to spatial data.

In the remainder of this chapter, we discuss three recently proposed multivariate event detection methods: the multivariate Bayesian scan statistic [Neill *et al.* (2007), Neill and Cooper (2008)], the agent-based Bayesian scan statistic [Jiang and Cooper (2008)], and the Anomalous Group Detection method [Das *et al.* (2008)]. All of these methods incorporate a Bayesian network structure to efficiently model the relationships between variables in the multivariate dataset, using the observed variables to draw inferences about which type of event has occurred and which subset of the data has been affected. The multivariate Bayesian scan statistic (MBSS) and agent-based Bayesian scan statistic (ABSS) methods each assume a fixed Bayesian network structure relating the underlying event to the observed variables and unobserved state variables, while the Anomalous Group Detection (AGD) method *learns* the Bayesian network structure from data. All three methods can be considered generalizations of the

simple Bayesian network anomaly detection method discussed above, detecting self-similar groups of anomalous records and characterizing the discovered patterns. They also generalize the use of scan statistics to detect clusters of counts, extending spatial scan methods from simple univariate models to multivariate datasets, and thus providing a general and powerful framework for event detection. AGD can also be applied to more general pattern detection problems which may not have a spatial or temporal structure, such as knowledge discovery from scientific databases.

1.2 The multivariate Bayesian scan statistic

The multivariate Bayesian scan statistic (MBSS) is a general framework for event detection and characterization using multivariate stream-based data. The MBSS method was first presented by Neill *et al.* (2007) and further developed by Neill and Cooper (2008). This approach extends the original, univariate Bayesian spatial scan statistic [Neill *et al.* (2006)] in two ways. First, rather than detecting patterns in a single stream of data, it integrates information from multiple data streams, improving the timeliness and accuracy of event detection. Second, MBSS extends the Bayesian framework to model and distinguish between multiple different types of events, thus enabling both detection and characterization of events.

In the stream-based event detection problem, we are given a dataset D consisting of multiple data streams D_m. Each data stream contains spatial time series data collected at a set of spatial locations s_i. For each stream D_m and location s_i, we have a time series of counts $c^t_{i,m}$, where $t = 0$ represents the current time step and $t = 1 \ldots T$ represent the counts from 1 to T time steps ago respectively. In disease surveillance, the data streams may include Emergency Department (ED) visits, with each stream representing the number of visits with a different chief complaint type, and over-the-counter (OTC) medication sales, with each stream representing the number of sales of a different product group. Thus a given count $c^t_{i,m}$ might represent the number of respiratory ED visits, or the number of cough/cold drugs sold, for zip code s_i on day t.

The goals of the MBSS method are event detection and characterization: to detect any relevant events occurring in the data, identify the type of event, and determine the event duration and affected locations. Thus MBSS compares the set of alternative hypotheses $H_1(S, E_k)$, each representing the occurrence of some event of type E_k in some space-time region S, against the null hypothesis H_0 that no events have occurred. In disease surveillance, the event types may be either specific illnesses (e.g. influenza, anthrax), non-specific syndromes (e.g. influenza-like illness), or other non-outbreak events that may result in

patterns of increased counts, such as promotional sales of OTC medications, inclement weather, or tourism. More generally, an event can be thought of as a process that affects some subset of the count data $c_{i,m}^t$ in some probabilistic manner. In addition to the set of event types E_k, MBSS is also given the set of space-time regions S to search, where each region S contains some subset of the counts $c_{i,m}^t$. Typically, each search region represents some set of spatial locations s_i for some time duration w, and regions of varying size, shape, and duration are considered.

1.2.1 Methods

Given the set of event types E_k, the set of space-time regions S, and the multivariate dataset D, MBSS computes the posterior probability $\Pr(H_1(S, E_k) \mid D)$ that each event type E_k has affected each space-time region S, as well as the posterior probability $\Pr(H_0 \mid D)$ that no event has occurred. The prior probability of each event type occurring in each space-time region, $\Pr(H_1(S, E_k))$, and the prior probability of no events, $\Pr(H_0)$, are given. MBSS computes the likelihood of the multivariate data given each hypothesis, and then calculates the posterior probability of each hypothesis using Bayes' Theorem:

$$\Pr(H_1(S, E_k) \mid D) = \frac{\Pr(D \mid H_1(S, E_k))\Pr(H_1(S, E_k))}{\Pr(D)} \tag{1.4}$$

$$\Pr(H_0 \mid D) = \frac{\Pr(D \mid H_0)\Pr(H_0)}{\Pr(D)} \tag{1.5}$$

Here the total probability of the data, $\Pr(D)$, is equal to $\Pr(D \mid H_0)\Pr(H_0) + \sum_{S,E_k} \Pr(D \mid H_1(S, E_k))\Pr(H_1(S, E_k))$.

In the MBSS framework, counts are assumed to have been generated from the Bayesian network represented in Figure 1.2. The event type E_k is drawn from a multinomial distribution: here $k = 0$ represents the null hypothesis H_0 of no events, with probability $\Pr(H_0)$, and $k = 1 \ldots K$ represent the occurrence of event type E_k, with corresponding probabilities $\Pr(E_k)$. The region of effect S is conditional on the event type, with probabilities $\Pr(H_1(S, E_k) \mid E_k)$. The distribution of event types and regions can be learned from training data or obtained from expert knowledge.

The effects of an event $H_1(S, E_k)$ on the data are determined by a value $x_{i,m}^t$ for each location s_i, data stream D_m, and time step t. These effects are assumed to be multiplicative, increasing the expected value of each count $c_{i,m}^t$ by a factor of $x_{i,m}^t$. For the null hypothesis H_0, no events have occurred, and $x_{i,m}^t = 1$ everywhere. For an event $H_1(S, E_k)$, only locations and time steps inside the space-time region S have been affected, and thus $x_{i,m}^t = 1$ for all $i, m, t \notin S$. Each event type can have a different joint probability distribution over the effects $x_{i,m}^t$.

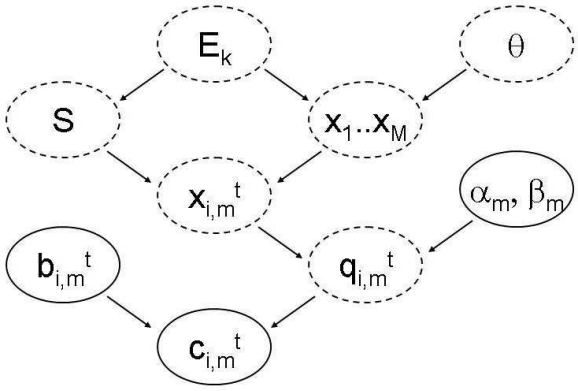

Figure 1.2: Bayesian network representation of the MBSS method. Solid ovals represent observed quantities, and dashed ovals represent hidden quantities that are modeled. The counts $c_{i,m}^t$ are directly observed, while the baselines $b_{i,m}^t$ and the parameter priors for each stream (α_m, β_m) are estimated from historical data.

The current implementation of MBSS [Neill and Cooper (2008)], as applied to the disease surveillance domain, makes several additional assumptions. To determine the search regions S, spatial locations are mapped to a uniform grid, and all gridded rectangular regions are considered. This method yields computational efficiency and the ability to detect both compact and elongated clusters [Neill *et al.* (2005a)]. MBSS assumes a hierarchical Gamma-Poisson model [Clayton and Kaldor (1987), Mollié (1999)]: each count $c_{i,m}^t$ is drawn from a Poisson distribution with mean proportional to the product of the expected count $b_{i,m}^t$ and the relative risk $q_{i,m}^t$. The expected counts (assuming no events taking place), are inferred from historical data, accounting for day-of-week and seasonal trends. Under the null, all relative risks $q_{i,m}^t$ for a given data stream D_m are drawn independently from a Gamma distribution with parameters (α_m, β_m). These parameters are estimated for each data stream by matching the mean and variance of the Gamma-Poisson model to their observed values in historical data. Under the alternative hypothesis $H_1(S, E_k)$, the relative risks $q_{i,m}^t$ inside region S are drawn from a Gamma distribution with parameters ($x_{i,m}^t \alpha_m$, β_m). Neill and Cooper (2008) assume a simplified event model, in which an event's effect on each data stream D_m is some constant x_m. These constants are a function of the average effects $x_{km,avg}$ of event type E_k on data stream D_m, as well as the event severity θ: $x_m = 1 + \theta(x_{km,avg} - 1)$. For example, consider an event type E_k with average effects $x_{km,avg} = 1.5$, 1.2, and

1.0 on three data streams $D_1 \ldots D_3$. For an event of "average" severity ($\theta = 1$), the expected counts of streams D_1 and D_2 would be increased by 50% and 20% respectively, with no effect on stream D_3. For a more severe event with severity $\theta = 2$, the expected counts of streams D_1 and D_2 would be increased by 100% and 40% respectively. Neill and Cooper (2008) assume a fixed, discrete distribution for θ, and present a simple, smoothed maximum likelihood method for learning the average effects $x_{km,avg}$ from labeled training examples.

The marginal likelihood of each observed count $c^t_{i,m}$ can be computed given the effect $x^t_{i,m}$, baseline $b^t_{i,m}$, and parameter priors α_m and β_m. MBSS integrates over all possible values of the relative risk $q^t_{i,m}$, weighted by their respective probabilities. Neill and Cooper (2008) derive a closed form (negative binomial) solution for the marginal likelihood. Since the null hypothesis assumes $x^t_{i,m} = 1$ everywhere, and since the counts are conditionally independent given the baselines, the α and β parameters, and the effects $x^t_{i,m}$, the marginal likelihood of the data under the null hypothesis can be easily computed. To calculate the likelihood of the data given an alternative hypothesis $H_1(S, E_k)$, MBSS marginalizes over the distribution of effects $x^t_{i,m}$, computing a weighted average of the data likelihoods given each effects vector $(x_1 \ldots x_M)$, weighted by the conditional probability of those effects given $H_1(S, E_k)$. The simplified event model makes these marginals efficiently computable: for each possible event type and severity, MBSS computes log-likelihood ratios for each location, and then computes the log-likelihood ratios for all regions under consideration by summing the location log-likelihoods. Alternatively, we can efficiently find those regions with highest log-likelihood ratios, using a variant of the fast spatial scan [Neill and Moore (2004)].

1.2.2 Evaluation

Neill and Cooper (2008) evaluated the event detection and characterization performance of the MBSS method, with and without incorporating prior information, on simulated outbreaks of influenza-like illness (ILI) injected into three streams of over-the-counter medication sales data (cough/cold, antifever, and thermometers) from Allegheny County, Pennsylvania. A "general" MBSS detector was used to handle the case when no prior knowledge of events is available. This detector assumed $2^M - 1$ event models (one for each non-empty subset of the M data streams). Each event model assumed equal average effects on the affected subset of streams, and assumed a uniform prior over the event types and affected regions. A "specific" MBSS detector was used to handle the case when prior knowledge of one or more event types is available. This detector assumed a pre-specified event model for each event type, giving the average effects of this event type on each data stream. The main results of their evaluation include:

1. The "general" MBSS detector achieved 1.5 days faster detection than univariate BSS detectors monitoring each data stream separately, demonstrating that MBSS increases detection power by integrating information from the multiple data streams.

2. The "general" MBSS detector and Kulldorff's multivariate spatial scan statistic [Kulldorff *et al.* (2007)] achieve very similar detection performance, suggesting that either method can be used to detect a broad range of event types when no prior information is available.

3. The "specific" MBSS detector was able to detect outbreaks an average of 1.3 days faster than either the "general" MBSS detector or Kulldorff's multivariate scan. This demonstrates that MBSS can achieve higher detection power by incorporating information about an event's effects on the different data streams. Further performance gains result from using informative region priors that incorporate knowledge of the distribution of each event type in space and time [Neill (2007)].

4. Given an event model for each of two different outbreak types (one primarily causing respiratory symptoms, and one primarily causing fever), MBSS was able to accurately differentiate between the outbreaks by the second outbreak day. The posterior probability of the correct outbreak type increased rapidly over the course of the outbreak, while the probability of the incorrect outbreak type remained constant and small.

1.2.3 Discussion

Neill and Cooper (2008) demonstrate that the MBSS method has several advantages as compared to prior event detection approaches. As in the univariate Bayesian spatial scan method [Neill *et al.* (2006)], MBSS can incorporate prior information of an event's effects and its distribution in space and time, increasing detection power. Similarly, the Bayesian scan statistics do not require randomization testing, resulting in 2-3 orders of magnitude faster computation as compared to the standard frequentist spatial scan.

Extension of the Bayesian framework to the multivariate case has further, substantial benefits. Integration of information from multiple data streams enables MBSS to detect emerging patterns (e.g. the early stages of an emerging outbreak of disease) that would not be visible from monitoring only a single stream. Incorporating multiple event models not only increases detection power, but also allows MBSS to *characterize* events by specifying models for multiple event types and computing the probability that each type of event has occurred. This enables the user to distinguish relevant events requiring urgent responses from irrelevant events which can safely be ignored, as well as informing the user's response to these events. For example, patterns of influenza-like

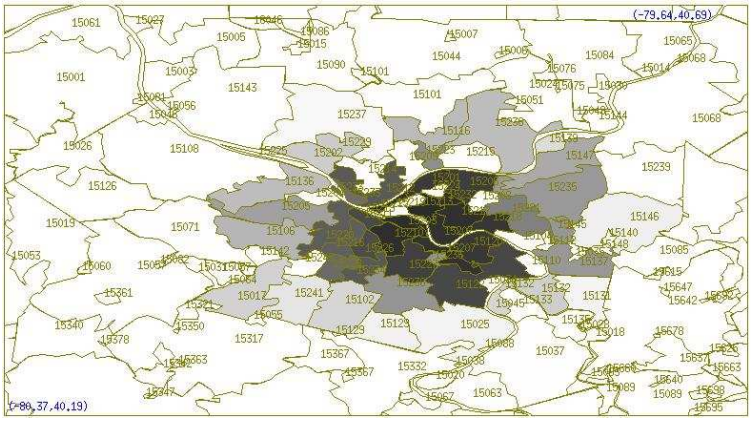

Figure 1.3: Example of a probability map computed by MBSS, from Neill and Cooper (2008). Darker shading indicates a higher probability that the given zip code has been affected.

illness would be a high priority for public health officials if these cases were due to pandemic avian influenza or a bioterrorist anthrax attack, and different interventions would be necessary in each case.

Finally, the outputs of MBSS (posterior probabilities of each event type in each space-time region) are easy to interpret, visualize, and use for decision-making. For example, considering the posterior probabilities of a given event type E_k on a given day t, we can compute the probability that each spatial location has been affected by summing the probabilities of all regions containing that location, and display the resulting "probability map" (Figure 1.3).

Comparison to prior methods

The Bayesian network shown in Figure 1.2 is a special case of the general stream-based scan statistic in Figure 1.4. In the general case, the counts $c_{i,m}^t$ are conditionally independent given the baselines $b_{i,m}^t$ and relative risks $q_{i,m}^t$. The joint distribution of the $q_{i,m}^t$ is conditional on the event type E_k and region S. However, the values of $q_{i,m}^t$ (for each location s_i, stream D_m, and time step t) may be correlated by dependence on other hidden nodes. For example, in Figure 1.2, observing a stream with a high count makes it more likely that the event severity θ is large, and thus increases the probability that another stream has a high count.

Both the univariate Bayesian spatial scan statistic [Neill *et al.* (2006)] and Kulldorff's Poisson spatial scan statistic [Kulldorff (1997)] can be considered special cases of the Bayesian network in Figure 1.4, assuming a single data stream D_m and a single event type ($E_k = H_1$ or H_0). In either case, we assume

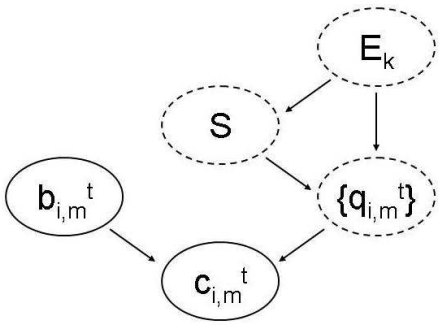

Figure 1.4: General Bayesian network representation of stream-based scan approaches. Relative risks $q^t_{i,m}$ are conditioned on the event type E_k and region S, and may be correlated. Counts $c^t_{i,m}$ are conditionally independent given the relative risks $q^t_{i,m}$ and baselines $b^t_{i,m}$.

three additional nodes in the Bayesian network (q_{in}, q_{out}, q_{all}). Under the null hypothesis H_0, $q^t_{i,m} = q_{all}$ everywhere, and under the alternative hypothesis $H_1(S)$, $q^t_{i,m} = q_{in}$ inside region S and $q^t_{i,m} = q_{out}$ outside region S. Kulldorff's Poisson scan statistic assumes the maximum likelihood values for q_{in}, q_{out}, and q_{all}. The Bayesian spatial scan statistic instead marginalizes over each value, assuming that $q_{in} \sim \text{Gamma}(x_{in}\alpha_{in}, \beta_{in})$, $q_{out} \sim \text{Gamma}(\alpha_{out}, \beta_{out})$, and $q_{all} \sim \text{Gamma}(\alpha_{all}, \beta_{all})$. The values of the α and β parameters are learned from data, and a discrete uniform distribution of x_{in} is assumed.

We note that the MBSS model differs from the original univariate Bayesian spatial scan model [Neill *et al.* (2006)] even for the case of a single data stream and single event type. Like Kulldorff's spatial scan statistic [Kulldorff (1997)], the original Bayesian spatial scan assumes constant relative risks q_{in}, q_{out}, and q_{all}. The MBSS model allows these risks to vary over space, time, and for different data streams, assuming that each risk is drawn independently from the Gamma distribution for that stream. Allowing risks to vary under the null hypothesis reduces the number of false positives due to overdispersion of counts, and the MBSS framework defines a simple and efficiently computable model for the impact of each event type on each data stream.

Incorporating learning into pattern detection

One important aspect of MBSS is the ability to learn new event models (and incrementally update existing models) from user feedback or from labeled training data. Neill and Cooper (2008) demonstrate that the average effects of each

event type can be learned from a small number of labeled examples, and that the fitted models gained a large improvement (average of 1.3 days faster detection) as compared to the general multivariate detectors. We note that learning from data may only be feasible for very common outbreaks (e.g. influenza), while models of rare events would still rely heavily on expert knowledge. Another possibility would be to learn models of common "confounding" events which are not relevant for detection, and use these models to reduce the false positive rates. For example, patterns of over-the-counter sales of cough/cold medications may occur due to cold weather, poor air quality, short-term population fluctuations due to tourism, or even promotional sales of these medications.

Future work

The incorporation of incremental model learning into the multivariate Bayesian pattern detection framework will be an important aspect of future work. In addition to the effects of each event type on the multiple data streams, many other aspects of the event models can be learned from labeled data, including the prevalence, size, shape, and spread of each type of event. The preliminary results of Neill (2007) suggest that learning these aspects of the event model can also lead to significant improvements in detection performance. Additionally, "active learning" methods can be incorporated in order to choose potential events that are both most relevant to the user and most informative to the system, present these events to the user, and update models based on the user feedback. Finally, the current MBSS implementation assumes the occurrence of a single event, with constant effects over time. Future work will include extending MBSS to "dynamic models" (where events can move and grow over time, and can have spatially and temporally varying effects), as well as "synergistic models" (where multiple events with interacting effects can occur).

1.3 The agent-based Bayesian scan statistic

Most existing approaches to event detection are "stream-based" methods which monitor the aggregate counts of a set of data streams and report patterns of anomalously high counts. For example, a stream-based disease surveillance system such as MBSS may look at the daily sales of anti-diarrheal medication and numbers of gastrointestinal ED visits, with the goal of detecting an outbreak of *Cryptosporidium*. An alternative event detection approach is to model each individual (agent) in a population, observe one or more variables for each individual, and draw inferences about the underlying event. These "agent-based" approaches often rely on an explicit Bayesian network representation to model

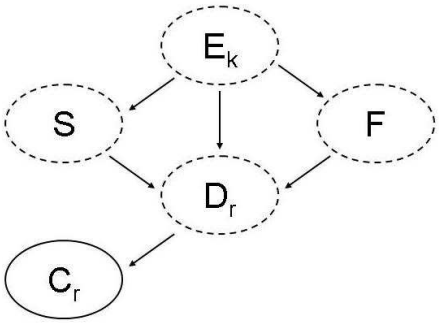

Figure 1.5: Bayesian network representation of the ABSS method. Solid ovals represent observed quantities, and dashed ovals represent hidden quantities that are modeled. Each agent's value of C_r is directly observed.

the causal relationships between the underlying event, the state of each individual (which usually cannot be directly observed), and the observable variables. The PANDA system developed by Cooper *et al.* (2004, 2007) is an agent-based Bayesian network approach for disease surveillance using Emergency Department data. Here we consider the agent-based Bayesian scan statistic (ABSS) method proposed by Jiang and Cooper (2008), which extends PANDA by incorporating spatial information.

The ABSS approach assumes a population of R agents, $r = 1 \ldots R$. Each agent might represent an individual in the population, a measurement device (e.g. a sensor that monitors for the presence of microbes), or some other entity. Each agent r has a set of observable values C_r, which is conditioned on that agent's underlying state D_r. As in Jiang and Cooper (2008), we assume here that agents are individuals in the population, and that each individual has a single observable value C_r drawn from some multi-valued discrete distribution. For example, in the disease surveillance domain, D_r may represent an individual's underlying disease state, which is not directly observed, and C_r may represent that individual's Emergency Department (ED) visit or purchase of over-the-counter (OTC) medication. The underlying states, and therefore the observable values, are conditioned on the event type E_k and the affected region S, enabling us to draw inferences about the event and affected region given the set of observed values $\{C_r\}$.

1.3.1 Methods

As in the MBSS approach, the agent-based Bayesian scan statistic assumes a fixed set of event types E_k and a fixed set of spatial regions S. Given the multi-

variate dataset D, the goal of this method is to compute the posterior probability $\Pr(H_1(S, E_k) \mid D)$ that each event type has occurred in each spatial region, as well as the posterior probability $\Pr(H_0 \mid D)$ that no events have occurred. These probabilities can be computed by Bayes' Theorem (Equation 1.4), combining the prior probability of each hypothesis with the data likelihood given that hypothesis.

However, the agent-based approach, rather than being given spatial time series data, is given a value C_r for each individual in the population, $r = 1 \ldots R$. These values are assumed to be drawn from some multi-valued discrete distribution, and are conditionally independent of other individuals' values given the individual's underlying state D_r (drawn from a different multi-valued discrete distribution). As shown in the Bayesian network representation in Figure 1.5, each individual's state D_r is conditionally independent given the event type E_k, the spatial region of effect S, and the fraction F of the population that has been affected.

Jiang and Cooper (2008) apply their agent-based approach to the detection of disease outbreaks using Emergency Department chief complaint data. The chosen Bayesian network representation is an extension of the Bayesian network used in PANDA-CDCA [Cooper *et al.* (2007)]. PANDA-CDCA does not incorporate spatial information, but ABSS adds an extra node to the Bayesian network representing the spatial region of effect S. In the ABSS framework, the event type E_k is assumed to take on one of 14 values: the 13 different outbreak diseases considered in PANDA-CDCA (influenza, anthrax, etc.) or H_0 (no outbreak occurring). Each individual's underlying state D_r represents two quantities: whether or not the individual goes to the Emergency Department, and in the event of an Emergency Department visit, what disease is responsible for the visit. Thus D_r can take on 15 different values: the 13 different outbreak types, "other" (i.e. the individual goes to the ED for another reason, such as an accident or broken bone), or "no ED" (i.e. the individual does not visit the Emergency Department). The observed values C_r represent the chief complaints for each ED patient (or "no ED" for individuals who did not visit the ED). As in PANDA-CDCA, chief complaints were classified into 54 different categories, and thus each C_r can take on 55 different values including "no ED".

In Jiang and Cooper (2008), the conditional probability table for each node of the Bayesian network in Figure 1.5 is pre-specified based on expert knowledge of the domain. The prior distribution $\Pr(E_k)$ assumes $\Pr(H_0) = 0.95$, $\Pr(\text{influenza}) = 0.04$, and small priors on the 12 other outbreak types (for example, $\Pr(\text{botulism}) = 0.0005$). As in MBSS, the events are assumed to be mutually exclusive, and thus $\Pr(H_0) + \sum_k \Pr(E_k) = 1$. Each event type E_k is assumed to have a uniform region prior, $\Pr(H_1(S, E_k) \mid E_k) = \frac{1}{N_{regions}}$, where $N_{regions}$ is the total number of spatial regions considered. More generally, each event type could have a different spatial prior distribution over regions, and

these distributions could be either pre-specified by expert knowledge or learned from labeled training data (e.g. known outbreaks). The variable F is assumed to represent the fraction of the population that is affected by the outbreak and goes to the ED. In the current implementation of ABSS, Jiang and Cooper (2008) assume a fixed, discrete distribution for F. However, different outbreak types might tend to affect different fractions of the population, or be more or less likely to send affected individuals to the ED. The dependence of F on the event type E_k in Figure 1.5 allows this information to be incorporated as well.

The distribution of D_r depends on whether any outbreak is occurring, and if so, whether individual r is in the affected spatial region S. In the event of no outbreak, or for individuals outside S, D_r is assigned the values "other" or "no ED", where the probability of an individual visiting the ED is estimated using historical data. For individuals inside region S when an outbreak is occurring, D_r is assigned either the outbreak disease (with probability F), "other", or "no ED". Finally, each outbreak disease D_r (including "other") has its own probability distribution over chief complaints C_r, and these distributions were specified by a domain expert.

We now consider how to compute the likelihood of the data for a given event type E_k, affected region S, and fraction F. For the null hypothesis H_0, the same inference can be performed, assuming that $S = \emptyset$. Given the observed value C_r for each individual $r = 1 \ldots R$, Jiang and Cooper (2008) perform inference on the Bayesian network, marginalizing over the values of the hidden nodes D_r:

$$\Pr(D \mid H_1(S, E_k, F)) = \prod_r \sum_{D_r} \Pr(C_r \mid D_r) \Pr(D_r \mid H_1(S, E_k, F))$$

$$= \prod_{r \in S} \sum_{D_r} \Pr(C_r \mid D_r) \Pr(D_r \mid H_1(E_k, F)) \times \prod_{r \notin S} \sum_{D_r} \Pr(C_r \mid D_r) \Pr(D_r \mid H_0)$$

The total likelihood of the data given each hypothesis can be calculated by marginalizing over the distribution of F, and the posterior probabilities can be computed from the likelihoods and priors using Bayes' Theorem as above.

1.3.2 Evaluation

Jiang and Cooper (2008) evaluated ABSS on simulated outbreaks of influenza and cryptosporidium, injected into real-world Emergency Department data from Allegheny County, Pennsylvania. Detection power (average days to detection, as a function of the allowable false positive rate) and spatial detection accuracy (average overlap between true and detected clusters) were compared to two previously proposed methods, PANDA-CDCA [Cooper *et al.* (2007)] and Kulldorff's original (univariate) spatial scan statistic [Kulldorff (1997)]. Their comparisons demonstrate that ABSS outperformed both PANDA-CDCA and spatial scan by a substantial margin for both datasets and according to both

performance measures. The improvement over PANDA-CDCA, which does not use spatial information, demonstrates that incorporation of spatial information into the agent-based Bayesian network framework substantially improves detection power. The improvement over spatial scan, which only uses the aggregate case count in each spatial area rather than the counts for each individual symptom, demonstrates that incorporation of multivariate information (and modeling of the underlying causal structure) also enables improved detection.

1.3.3 Discussion

The agent-based Bayesian scan statistic model can be considered a variant of standard scan statistic approaches where data is provided for each individual in the population rather than for a set of data streams. This model is particularly appropriate when we have individual-level data, but can be used for aggregate count data as well. Using individual-level data, if this data is available, has several advantages. Though the current ABSS model assumes that each individual r has the same probability distribution for their underlying state D_r and observed variable C_r, the model can be easily extended to the case where these distributions are conditioned on individual-level covariates such as age, gender, and occupation. Additionally, the agent-based model can be extended to the case where each individual has a joint distribution over multiple observable variables. Observing a single individual with multiple indicators of an event (for example, an ED patient who has both a fever and a rash) may enable faster and more accurate detection than separately considering the number of individuals with each indicator.

On the other hand, if only the aggregate counts are provided, then either the agent-based (multinomial) or the stream-based (multivariate Poisson) method may be more appropriate. For example, we may observe only the number of ED patients with each chief complaint type, or the total sales of each category of over-the-counter medication. If the number of individuals in the population is known, and each individual can take only one action (such as visiting the ED with a specific chief complaint type) out of a predefined set of actions, then the ABSS model may be preferable. If individuals can take multiple actions, and the population size is not known, we might prefer to infer the expected counts from historical data and compare actual to expected counts, as in MBSS.

Comparison to prior methods

The Bayesian network shown in Figure 1.5 is a special case of the general agent-based scan statistic in Figure 1.6. In the general case, each individual r has an observed value C_r. The joint distribution of the C_r is conditional on the event type E_k and region S. However, different individuals' values of C_r may be correlated by the addition of hidden nodes to the Bayesian network. For

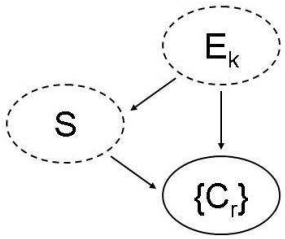

Figure 1.6: General Bayesian network representation of agent-based scan approaches. Solid ovals represent observed quantities, and dashed ovals represent hidden quantities that are modeled. Each agent's value of C_r is conditioned on the event type E_k and region S, and these values may be correlated by additional hidden nodes.

example, in Figure 1.5, observing an individual with disease symptoms increases the likelihood that F (the fraction of the population affected) is large, and thus increases the probability that another individual has disease symptoms.

The Bernoulli spatial scan statistic [Kulldorff (1997)] can also be considered a special case of the Bayesian network in Figure 1.6, with one event type ($E_k = H_1$ or H_0) and binary variables C_r. In this case, we assume three additional nodes in the Bayesian network (q_{in}, q_{out}, q_{all}). Under the null hypothesis H_0, $\Pr(C_r = 1) = q_{all}$ everywhere, and under the alternative hypothesis $H_1(S)$, $\Pr(C_r = 1) = q_{in}$ inside region S and $\Pr(C_r = 1) = q_{out}$ outside region S. However, rather than marginalizing over q_{in}, q_{out}, and q_{all}, the Bernoulli spatial scan assumes the maximum likelihood values for each node.

Future work

Future work by Jiang et al. will compare the agent-based approach to other multivariate spatial detection methods, including MBSS and Kulldorff's multivariate spatial scan statistic. Additionally, the current implementations of ABSS and PANDA-CDCA used a "specific" detector with pre-specified models of 13 outbreak diseases (including influenza, cryptosporidium, and the CDC Category A diseases), and the simulated outbreaks were generated assuming a distribution of chief complaints that is identical to these models. Future work will evaluate ABSS on disease outbreaks generated according to different chief complaint distributions (i.e. measuring performance as a function of the difference between true and assumed distributions), and thus test the robustness of this method to model misspecification. While the current implementation of ABSS is specific to Emergency Department disease surveillance, ABSS can

be extended to other application domains using more general definitions of the individual's underlying state D_r, observed behavior C_r, and the fraction of the population affected F. Finally, future versions of ABSS will include many of the current features of MBSS, such as incorporation of temporal information, visualization of outputs, and learning of event models from labeled data.

1.4 The Anomalous Group Detection method

We now consider how the scan statistic framework can be extended from the specific case of event surveillance to more general multivariate datasets. This extension poses several challenges. Since many datasets have no explicit space or time component, we cannot simply search over geographical regions, and thus it is not clear which subsets of the data should be considered. Additionally, while other scan statistic methods assume a fixed parametric model for the effects of different types of pattern on the data, we may wish to detect anomalous patterns in more general datasets where no such model is known. One solution to these challenges is provided by the Anomalous Group Detection (AGD) method, recently proposed by Das *et al.* (2008). Rather than relying on a fixed parametric model, AGD *learns* the structure and parameters of a Bayesian network from the data, and searches over self-similar subsets of the data to find anomalous patterns.

The AGD method can be used to detect anomalous groups in arbitrary, non-spatial datasets with discrete valued attributes. For typical stream-based scan statistic approaches, each data point consists of a set of real-valued "location" attributes as well as real-valued "count" data. The set of search regions is defined by the location attributes (e.g. spatial scan searches over geographically contiguous subsets of the data) while the likelihood under each hypothesis $H_1(S)$ is a function of the counts inside and outside region S. In the more general pattern detection problem, there may be no defined set of location attributes, and thus we can no longer predefine a set of search regions based on geographical attributes such as size, shape, or contiguity. Nevertheless, we want to formulate a measure of how well a subset of data points fit as a *group* based on the similarity between them. We must then perform a search over all possible subsets of the data in order to find the most anomalous groups.

Another difference between the AGD method and other scan statistic approaches is in the definition of anomalousness for a data point or a group of points. Scan statistics are usually applied to detect overdensities of records in a given space: individual records are aggregated into counts, and clusters with anomalously high counts are detected. In the AGD framework, however, each record has many discrete-valued attributes, and can have an inherent degree of

anomalousness depending on its features. Most records are generated from the "normal" distribution of data and hence are not relevant. Instead, the goal of AGD is to detect groups of records that are both anomalous and also self-similar in some respect.

1.4.1 Methods

The AGD framework assumes a multivariate dataset D, where each data record $R_i \in D$ has values for a set of discrete-valued attributes $X_1 \ldots X_M$. As in the original spatial scan statistic approach [Kulldorff (1997)], AGD finds the set of records that maximizes the likelihood ratio statistic $F(S) = \frac{\Pr(D \mid H_1(S))}{\Pr(D \mid H_0)}$, where H_0 is the null hypothesis that there are no anomalies present, and $H_1(S)$ is the alternative hypothesis specifying that the set S is an anomalous group. AGD assumes Bayesian network models for both the null and alternative hypothesis, and computes the data likelihoods given these models. For the null hypothesis H_0, a Bayesian network model is inferred from a separate training dataset (e.g. historical data), which is assumed to contain no anomalies, and all data records are assumed to have been drawn independently from this model. Under the alternative hypothesis $H_1(S)$, the records contained in subset S are assumed to have been drawn from a different Bayes Net model, while the rest of the data records are generated from the null model. The Bayesian network model parameters for the alternative hypothesis $H_1(S)$ are learned directly from the records in subset S, as discussed below.

This scoring metric gives a higher score to anomalous records, as well as setting a constraint of similarity between the records in a group. If the records in S are similar to each other, then $H_1(S)$ will be able to model them tightly. This will result in a high value of the data likelihood under the alternative hypothesis $H_1(S)$, thus increasing the score $F(S)$. Also, records that are poorly modeled by the training data will have low likelihoods under the null hypothesis H_0, again increasing the group score $F(S)$. Hence maximizing this score leads to grouping of similar records and at the same time it prefers records that are anomalous (i.e. that have low likelihoods under the null hypothesis).

As discussed by Das *et al.* (2008), the AGD algorithm consists of three steps:

1. Learn the Bayesian Network model for the null hypothesis H_0 from the training data.
2. For all subsets of the data S:
   (a) Fit the alternate hypothesis Bayesian Network ($H_1(S)$) parameters using data from subset S.
   (b) Compute the group likelihood ratio score $F(S)$.
3. Output the groups with highest score.

Step 1 is to learn the Bayesian network corresponding to the null hypothesis. The network structure is learned automatically from the training dataset using the Optimal Reinsertion algorithm [Moore and Wong (2003)], and this structure is assumed for the null hypothesis H_0 and for all alternative hypotheses $H_1(S)$. The probability table parameters of H_0 are then learned from the training dataset using smoothed maximum likelihood estimation. For a given node corresponding to the variable X_i in the Bayes Net, let X_{Π_i} denote the set of variables corresponding to the parent nodes of X_i. The conditional probability table of X_i has parameters corresponding to the conditional probability values $\theta_{ijk} = \Pr(X_i = j \,|\, X_{\Pi_i} = k)$. Here we must estimate θ_{ijk} for each variable X_i, value j, and set of parent values k. The maximum likelihood parameter estimates are given by $\hat{\theta}_{ijk} = \frac{N_{ijk}}{\sum_{j'} N_{ij'k}}$, where N_{ijk} denotes the number of instances in the training dataset with $X_i = j$ and $X_{\Pi_i} = k$. To deal with sparsity of the training data, Das *et al.* (2008) apply Laplace smoothing to adjust the estimate of each model parameter.

Steps 2-3 find groups of records S that maximize the likelihood ratio score $F(S) = \frac{\Pr(D \,|\, H_1(S))}{\Pr(D \,|\, H_0)}$, where the alternative hypothesis $H_1(S)$ assumes that the records in subset S form an anomalous group, and the null hypothesis H_0 assumes that no anomalous groups are present. The optimal group can be found by searching over all subsets of the test data, but this exhaustive search would require exponential time. Thus Das *et al.* (2008) propose a greedy heuristic search method which starts from each record as an initial seed and iteratively adds the record that most improves the likelihood ratio score. This search method can find high-scoring groups in a computationally efficient manner, but does not guarantee that the optimal group will be found.

Step 2a fits the parameters of the Bayesian network for the alternative hypothesis $H_1(S)$. Das *et al.* (2008) use an empirical Bayes approach in which these parameters are estimated from the counts in the subset of the test dataset represented by S, following an approach of smoothed maximum likelihood estimation similar to Step 1 above. In this case, N_{ijk} denotes the corresponding counts in region S. Since the number of records in group S may be small and this data is used to fit a large number of Bayesian network parameters, data sparsity is a serious problem, and computing the likelihood $\Pr(D \,|\, H_1(S))$ using this model risks overfitting of the data.

Step 2b computes the group likelihood ratio score $F(S)$, performing inference on the Bayesian Networks corresponding to $H_1(S)$ and H_0 to compute the data likelihoods under each hypothesis. Since data points are assumed to be conditionally independent given the model, and records not contained in subset S have identical likelihoods given $H_1(S)$ and H_0, the likelihood ratio statistic simplifies to:

$$F(S) = \frac{\prod_{R_i \in S} \Pr(R_i \,|\, H_1(S))}{\prod_{R_i \in S} \Pr(R_i \,|\, H_0)} \tag{1.6}$$

Das *et al.* (2008) deal with the overfitting problem mentioned above by using a "leave-one-out" method based on the pseudo-likelihood of each record R_i in S. In this case, the numerator of Equation 1.6 becomes $\prod_{R_i \in S} \Pr(R_i \mid H_1(S - \{R_i\}))$. To compute the likelihood of each record R_i, assuming the alternative hypothesis $H_1(S)$, a Bayesian network model is learned from all the records in S except for R_i, and this model is used to compute the likelihood of R_i. Since the likelihood of each record is computed without using that record to estimate the model parameters, this reduces the risk of over-fitting.

Step 3 outputs the highest scoring groups found in step 2. Additionally, Das *et al.* (2008) compute an anomalousness score for each individual record R in the test data, by finding the highest scoring group $S^*(R)$ that contains R. The score of record R can then be computed in one of two ways. In the "group likelihood ratio" approach, $Score(R)$ is set equal to the group score $F(S^*(R))$. This approach gives a high score to any record that is contained in a highly anomalous group, regardless of whether the record is itself anomalous or just similar to other anomalous records. Alternatively, we can consider only the contribution of record R to the score of $S^*(R)$. In this "single record likelihood ratio" approach, $Score(R)$ is set equal to the partial record pseudo-likelihood ratio, $\frac{\Pr(R \mid H_1(S^*(R) - \{R\}))}{\Pr(R \mid H_0)}$.

1.4.2 Evaluation

Das *et al.* (2008) compare the performance of their method to the baseline method described above, which detects individual records with low likelihoods given the null Bayes Net model. Synthetic anomalies were injected into two real-world datasets: a dataset of Emergency Department (ED) records from Allegheny County, Pennsylvania, and the PIERS dataset of container shipping data. The former dataset contains records of patients visiting Allegheny County Emergency Departments. Each record consists of six categorical attributes (hospital id, prodrome, age decile, home zip code and chief complaint class), and the goal is to detect anomalous groups of records (e.g. spatial disease clusters, age/gender clusters, and increases in different symptom types) that correspond to emerging disease outbreaks. The second dataset consists of records describing containers imported into the country. Each record consists of 10 attributes: country of origin, departing and arriving ports, shipping line, shipper name, vessel name, commodity being shipped, and the size, weight, and value of the container. In this case, the goal is to detect anomalous groups of records corresponding to patterns of smuggling, terrorist activity, or other illicit shipments.

Das *et al.* (2008) evaluated the performance of the algorithms in two different ways. The first evaluation criterion was the ability of each algorithm to identify each individual anomaly correctly. Figure 1.7 plots the detection

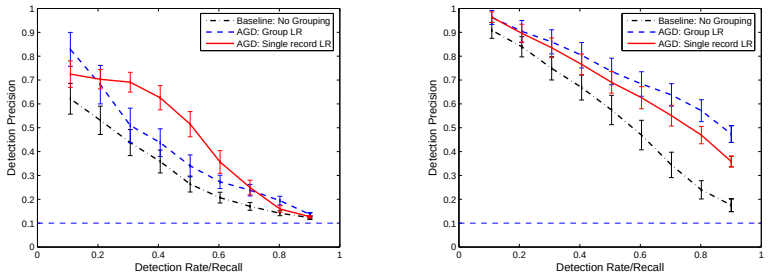

Figure 1.7: Plot of detection precision vs. recall for a) Emergency Department dataset and b) PIERS dataset, from Das et al. (2008).

precision, i.e. the ratio of number of true positives to the total number of predicted positives, against the detection rate, i.e. the proportion of total true anomalies that were detected. For both the "group likelihood ratio" and "single record likelihood ratio" methods, AGD performed significantly better than the baseline method without grouping. Similar results were obtained when examining the ability of the algorithms to identify and distinguish between entire datasets which have anomalous groups against ones which do not have any anomalies, e.g. distinguishing datasets containing outbreaks from datasets with no outbreaks. For these experiments, the grouping method again achieved significantly higher performance than the baseline anomaly detection method. While the set of anomalies were synthetically generated, current work by Das et al. includes evaluation on real anomalies, e.g. retrospective analysis of known disease outbreaks.

1.4.3 Discussion

The primary advantage of the AGD method is its generality: unlike the MBSS and ABSS methods, AGD can be directly applied to arbitrary multivariate datasets without the need for a pre-specified Bayesian network model of how the data is generated. Instead, the structure of the network and the parameters for each node are learned from a training dataset, and the learned model is used for detection. Although Das *et al.* (2008) exclusively deal with categorical valued datasets, AGD can be generalized to handle datasets containing real valued attributes as well, using Bayesian network models that incorporate both categorical and real valued nodes. However, AGD does have several disadvantages. It cannot model and distinguish between multiple event types, since the parameters for the alternative hypothesis $H_1(S)$ are fitted directly from that subset of the test data. Learning a model using the test data and then computing the likelihood of the test data given that model can result in overfitting, and the

proposed solution (use of the pseudo-likelihood) gives outputs that cannot be interpreted as posterior probabilities.

Comparison to prior methods

The AGD algorithm can be thought of as a generalization of scan statistic methods such as MBSS and ABSS to arbitrary multivariate datasets without predefined location or count attributes. All attributes of the data are used to determine both the self-similarity of the group and the anomalousness of its component records, as opposed to previous methods that determine the anomalousness of the count attributes and use the location attributes for grouping. While standard scan statistics implicitly or explicitly assume a fixed Bayesian network model relating the observed variables (i.e. aggregate counts in stream-based approaches, and individual-level variables in agent-based approaches) to the underlying event and affected region, AGD *learns* the underlying model from the training dataset. Additionally, standard scan statistics are geared toward the event detection problem, searching over a set of contiguous spatial regions that are predefined based on the location attributes of the data, while AGD performs a heuristic search over arbitrary subsets of the data.

Future work

Future work by Das et al. will extend the AGD approach to incorporate multiple pattern types E_k, model the effects of each pattern type on the data, and distinguish between multiple pattern types (by computing the posterior probability that each pattern type E_k affects each subset of the data S). Each pattern type can have a different prior probability $\Pr(E_k)$ and a different distribution over subsets of the data. Models of how each pattern type will affect a given subset of the data can be defined, allowing computation of the data likelihood given each hypothesis $H_1(S, E_k)$. Different pattern types can have a different distribution over Bayesian network structures and parameters, and the data can be represented as a "mixture of Bayes Nets." Each alternative Bayes Net model can be related to the null Bayes Net by changing the conditional distributions of the output attributes based on the event model E_k. Finally, future work will develop methods which can *learn* these models for each pattern type. These extensions could be valuable for finding groups in new datasets that match specific patterns of anomalous activity learned from earlier data.

Acknowledgements

This work was partially supported by NSF grant IIS-0325581 and CDC grant 8 R01 HK000020 02. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NSF or CDC.

References

Bronstein, A., Das, J., Duro, M., Friedrich, R., Kleyner, G., Mueller, M., Singhal, S., and Cohen, I. (2001). Bayesian networks for detecting anomalies in internet-based services. In *Intl. Symposium on Integrated Network Mgmt.*

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

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 *Proc. Conference on Uncertainty in Artificial Intelligence*.

Cooper, G. F., Dowling, J. N., Levander, J. D., and Sutovsky, P. (2007). A Bayesian algorithm for detecting CDC Category A outbreak diseases from emergency department chief complaints. *Advances in Disease Surveillance*, **2**, 45.

Das, K., Schneider, J., and Neill, D. B. (2008). Detecting anomalous groups in categorical datasets. Tech. rep. submitted for publication, Carnegie Mellon University, School of Computer Science.

Dong-Her, S., Hsiu-Sen, C., Chun-Yuan, C., and Lin, B. (2004). Internet security: malicious e-mails detection and protection. *Industrial Mgmt. and Data Sys.*, **104**, 613 – 623.

Duczmal, L. and Assuncao, R. (2004). A simulated annealing strategy for the detection of arbitrary shaped spatial clusters. *Computational Statistics and Data Analysis*, **45**, 269–286.

Heckerman, D., Geiger, D., and Chickering, M. (1995). Learning bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, **20**, 197–243.

Hjalmars, U., Kulldorff, M., Gustafsson, G., and Nagarwalla, N. (1996). Childhood leukemia in Sweden: using GIS and a spatial scan statistic for cluster detection. *Statistics in Medicine*, **15**, 707–715.

Jiang, X. and Cooper, G. F. (2008). A Bayesian network spatial scan statistic. Tech. rep. submitted for publication, University of Pittsburgh, Department of Biomedical Informatics.

Kleinman, K., Abrams, A., Kulldorff, M., and Platt, R. (2005). A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, **133**(3), 409–419.

Kulldorff, M. (2001). Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society A*, **164**, 61–72.

Kulldorff, M., Athas, W., Feuer, E., Miller, B., and Key, C. (1998). Evaluating cluster alarms: a space-time scan statistic and cluster alarms in Los Alamos. *American Journal of Public Health*, **88**, 1377–1380.

Kulldorff, M., Feuer, E. J., Miller, B. A., and Freedman, L. S. (1997). Breast cancer clusters in the northeast United States: a geographic analysis. *American Journal of Epidemiology*, **146**(2), 161–170.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006). An elliptic spatial scan statistic. *Statistics in Medicine*, **25**, 3929–3943.

Kulldorff, M., Mostashari, F., Duczmal, L., Yih, W. K., Kleinman, K., and Platt, R. (2007). Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, **26**, 1824–1833.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, **26**(6), 1481–1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: detection and inference. *Statistics in Medicine*, **14**, 799–810.

Mollié, A. (1999). Bayesian and empirical Bayes approaches to disease mapping. In Lawson, A. B., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J.-F., and Bertollini, R. (Eds.), *Disease Mapping and Risk Assessment for Public Health*.

Moore, A. and Wong, W.-K. (2003). Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. In *Proceedings of the 20th Intl. Conf. on Machine Learning*, pp. 552–559.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R., and Kulasekera, V. (2003). Dead bird clustering: A potential early warning system for West Nile virus activity. *Emerging Infectious Diseases*, **9**, 641–646.

Neill, D. B. (2006). Detection of spatial and spatio-temporal clusters. Tech. rep. CMU-CS-06-142, Ph.D. thesis, Carnegie Mellon University, School of Computer Science.

Neill, D. B. (2007). Incorporating learning into disease surveillance systems. *Advances in Disease Surveillance*, **4**, 107.

Neill, D. B. and Cooper, G. F. (2008). A multivariate Bayesian scan statistic for early event detection and characterization. Tech. rep. submitted for publication, Carnegie Mellon University, School of Computer Science.

Neill, D. B. and Lingwall, J. (2007). A nonparametric scan statistic for multivariate disease surveillance. *Advances in Disease Surveillance*, **4**, 106.

Neill, D. B., Moore, A. W., and Cooper, G. F. (2007). A multivariate Bayesian scan statistic. *Advances in Disease Surveillance*, **2**, 60.

Neill, D. B., Moore, A. W., and Sabhnani, M. R. (2005a). Detecting elongated disease clusters. *Morbidity and Mortality Weekly Report*, **54 (Supplement on Syndromic Surveillance)**, 197.

Neill, D. B., Moore, A. W., Sabhnani, M. R., and Daniel, K. (2005b). Detection of emerging space-time clusters. In *Proc. 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*.

Neill, D. B. and Sabhnani, M. R. (2007). A robust expectation-based spatial scan statistic. *Advances in Disease Surveillance*, **2**, 61.

Neill, D. B. and Moore, A. W. (2004). Rapid detection of significant spatial clusters. In *Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining*, pp. 256–265.

Neill, D. B., Moore, A. W., and Cooper, G. F. (2006). A Bayesian spatial scan statistic. In *Advances in Neural Information Processing Systems 18*, pp. 1003–1010.

Patil, G. P. and Taillie, C. (2004). Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Envir. Ecol. Stat.*, **11**, 183–197.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, **4**, 11.

Wong, W.-K., Moore, A. W., Cooper, G. F., and Wagner, M. M. (2003a). Bayesian network anomaly pattern detection for disease outbreaks. In *Proc. 20th International Conference on Machine Learning*.

Wong, W.-K., Moore, A. W., Cooper, G. F., and Wagner, M. M. (2003b). WSARE: What's strange about recent events?. *Journal of Urban Health*, **80**(2 Suppl. 1), i66–i75.

Ye, N. and Xu, M. (2000). Probabilistic networks with undirected links for anomaly detection. In *IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, pp. 175–179.