Scoring Function An objective function that measures the anomalousness of a subset of data

LTSS Linear-time subset scanning

Time to Detect Evaluation metric; time delay before detecting an event

Overlap Evaluation metric; accuracy of detected subsets of data

Detection Power Evaluation metric; proportion of detected events

Definition

GraphScan is a novel method for detecting arbitrarily shaped connected clusters in graph or network data. Given a graph structure, data observed at each node of the graph, and a score function defining the anomalousness of a set of nodes, GraphScan can efficiently and exactly identify the most anomalous (highest-scoring) connected subgraph. Additionally, GraphScan can be used to discover an unknown, underlying graph structure from unlabeled data.

Introduction

Many of the most interesting and relevant discoveries that can be made from data arise not from the evaluation of single records but from identifying a *group* of records that are collectively anomalous in some interesting way. To this end, the "subset scan" approach to pattern detection treats the problem as a search over subsets of data, with the goal of finding the most anomalous subsets. One major challenge of the subset scan approach is the computational problem that arises from attempting to search over the exponentially many subsets of the data. Linear-time subset scanning (LTSS) (Neill 2012) is a novel approach to anomalous pattern detection that addresses this issue by identifying the most anomalous subset of the data without requiring an exhaustive search, reducing computation time from years to milliseconds.

Social networks and many other data sources that contain emerging events of public interest

Disease Surveillance, Case Study

Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B. Neill
Event and Pattern Detection Laboratory, H. J. Heinz III College, Carnegie Mellon University, Pittsburgh, PA, USA

Synonyms

Biosurveillance; Event detection; Graph mining; Scan statistics; Spatial scan statistic

Glossary

Event Detection Identifying patterns of interest in large temporal datasets

Spatial Scan Statistic A method for identifying hotspots in spatial data, widely used in epidemiology and biosurveillance

are commonly represented with a graph structure. For such data, methods should identify the most anomalous subgraph or *connected subset* of records. Therefore, GraphScan (Speakman and Neill 2010; Speakman et al. 2012) was developed to extend the exactness and efficiency of the LTSS property to data sets that have an underlying graph structure.

In the disease surveillance domain, researchers are typically concerned with finding an anomalous spatial region which may be indicative of an emerging outbreak. Kulldorff's spatial scan statistic (Kulldorff 1997) can be used in these settings to detect circular clusters of anomalous locations. However, consider an outbreak from a waterborne illness that leads to an increased number of hospital visits from patients who live in zip codes along a river or coastline. This non-compact spatial pattern would be hard to detect using circular proximity constraints. Taking advantage of an underlying graph structure based on zip code adjacency allows GraphScan to consider sets of *connected* zip codes, increasing its power to detect these irregularly shaped clusters. Empirical results show that GraphScan is able to detect synthetic disease outbreaks several days earlier than the circular scan, with fewer than half as many missed outbreaks.

Historical Background

Anomaly detection in graphs can take on many different forms. The goal of GraphScan is to efficiently and exactly identify an anomalous subset of records that are connected to each other in the graph structure. It is not attempting to identify anomalous graph structure. In other words, the network structure is a *constraint* on the anomaly detection rather than the objective. For example, given a time series of the number of texts or calls of each individual in a friendship network, GraphScan can identify which group of friends (connected subgraph) is currently having the most anomalous activity. Both FlexScan (Tango and Takahashi 2005) and Upper Level Sets (ULS) (Patil and Taillie 2004) share this definition of anomalous subgraphs. However, FlexScan suffers from computational limitations, and ULS is not guaranteed to identify the most anomalous subgraph. These approaches are in contrast to (Akoglu et al. 2010), which identifies anomalous network structure within a node's egonet such as near-cliques or dominant heavy links between two nodes. AutoPart (Chakrabarti 2004) defines an information theoretic distance metric to identify subgraphs that are "far away" from other subgraphs as anomalous. Although the output of all these methods is subgraphs, the latter two methods measure anomalousness by edge weight or graph structure rather than activity at the nodes. From this point forward, we will reference anomalous subgraphs as anomalous subsets of records (nodes) that form a connected subgraph.

Basic Methodology

Spatial event detection methods typically monitor a data stream (such as Emergency Department visits with respiratory complaints or over-the-counter cough and cold medication sales) across a collection of spatial locations and over time. These streams are represented as a series of counts c_i^t, from location s_i, and time step t. This stream of counts is also used to determine the historical baselines (expected counts) b_i^t. The amount of anomalous activity in a region is quantified by the scoring function, $F(S) = F(C(S), B(S))$ where $C(S)$ and $B(S)$ represent the aggregate count and baseline of subset S, respectively.

In practice, these scoring functions are typically log-likelihood ratio statistics such as the expectation-based Poisson statistic: $F(S) = C \log(\frac{C}{B}) + B - C$, if $C > B$, and $F(S) = 0$ otherwise (Neill et al. 2005). The more that the total count exceeds the total baseline, the higher the score of the region. The goal of GraphScan is to *efficiently* and *exactly* identify the highest-scoring (i.e., most interesting) subset of data S, subject to the connectivity constraints of the underlying graph structure.

To do so, GraphScan builds on the "linear-time subset scanning" (LTSS) property (Neill 2012), a novel feature of commonly used scoring functions. For the expectation-based Poisson statistic, the highest-scoring subset of records can be found by first ordering the records according to their count-to-baseline ratio $\frac{c_i}{b_i}$. This ordering is referred to as a record's "priority" and represents the jth-highest priority record as $R_{(j)}$. For clarification, $R_{(1)}$ is the highest priority record. The highest-scoring subset can then be proven to consist of the top-j highest priority records for some (unknown) value of j. Please refer to Neill (2012) for more details on the LTSS property. In its simplest form, the LTSS property of scoring functions states: "If record $R_{(j)}$ is included in the optimal subset S, then all higher priority records $R_{(1)} \ldots R_{(j-1)}$ must also be included in that subset." When connectivity constraints are introduced, this statement of the LTSS property must be extended in order to be effectively applied to the task of identifying the highest scoring connected subset.

Enforcing Connectivity Constraints

There are two different reasons for including a record into a potentially highest-scoring connected subset. The first and most obvious reason is because the record contributes a large number of counts with a relatively low baseline: it is interesting of its own accord. However, some records may be included in the highest-scoring connected subset because they are simply enforcing the connectivity constraint, connecting other higher priority records while contributing very few counts and baselines themselves. Taking this observation into account, the GraphScan logic extends the LTSS property to state: "If $R_{(j)}$ is included in the highest scoring connected subset S, *and removing* $R_{(j)}$ *does not disconnect the subset*, then all higher priority neighbors of S must also be included in the optimal subset." Unlike the rule implemented within the ULS method, this approach is guaranteed to identify the highest-scoring connected subset because it allows lower priority records to be included in the optimal subset to enforce the connectivity constraints of the graph structure. In practice, this logic allows the GraphScan algorithm to prove that many connected subsets are suboptimal, excluding these subsets from the search without scoring each subset individually and thus dramatically reducing computation time. This efficient implementation of the LTSS property with additional connectivity constraints allows GraphScan to scale up to much larger data sets than FlexScan (Tango and Takahashi 2005), which naively scores all connected subsets of the graph structure. The details of the implementation of GraphScan are provided in Speakman et al. (2012).

Incorporating Proximity Constraints

The major focus of the GraphScan algorithm is combining *connectivity constraints* with the LTSS property in order to efficiently determine the highest-scoring connected subset of records. However, if the data set has both spatial and graph information available, then GraphScan may use both *proximity* and *connectivity* constraints simultaneously. For a given distance metric, a "local neighborhood" may be formed consisting of a central record and its $k-1$ nearest neighbors. This approach can be applied in the disease surveillance domain, where we have access to the latitude and longitude coordinates of the centroid of each zip code. For social networks, these additional constraints can be based on graph-distance (length of shortest path) between two nodes.

The GraphScan algorithm then finds the highest-scoring connected cluster *within* each of these "local neighborhoods" by forming a connectivity graph that only consists of the records in this neighborhood. The highest-scoring connected subset using this "k-nearest neighbor" approach is simply the highest-scoring connected subset found from the N possible neighborhoods.

Graph Structure Learning

If the underlying graph structure is known, we have seen that GraphScan (Speakman and Neill 2010) can be used to identify an anomalous subgraph, which may be indicative of an

emerging event. However, events might spread over some latent network structure, such as disease outbreaks spreading from person to person or information spreading through a social network. In Somanchi and Neill (2011, 2012), we present an approach for learning the network structure from unlabeled data, given only the time series of data at each network node.

Our solution builds on the GraphScan (Speakman and Neill 2010) and linear-time subset scanning (Neill 2012) approaches, comparing the most anomalous subsets detected with and without connectivity constraints. We consider a large set of potential graph structures; a greedy edge removal approach is used to search over the space of graph structures, as described in Somanchi and Neill (2012). We efficiently compute the highest-scoring connected subgraph for each graph structure and each training example using GraphScan. We normalize each score by dividing by the maximum unconstrained subset score for that training example (computed efficiently using LTSS). We then compute the mean normalized score averaged over all training examples. If a given graph is close to the true underlying structure, then its maximum graph-constrained score will be close to the maximum unconstrained score for many training examples. If the graph is missing essential connections, then the maximum graph-constrained score given that structure will be much lower than the maximum unconstrained score. However, any graph with a large number of edges will also score close to the maximum unconstrained score. Thus, we compare the mean normalized score of a given graph structure to the distribution of mean normalized scores for random graphs with the same number of edges, and we choose the graph structure with the most statistically significant score given this distribution.

Evaluation on Spatial Disease Surveillance

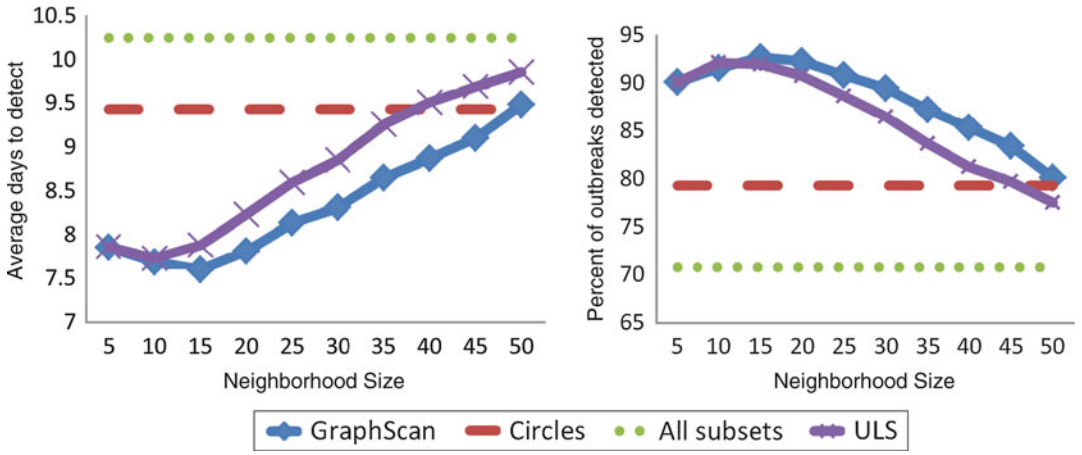
GraphScan's detection power was evaluated using a set of simulated respiratory disease outbreaks injected into real-world Emergency Department data from Allegheny County, Pennsylvania. Multiple methods were compared: "circles" (Kulldorff 1997) (traditional approach, returns the highest-scoring circular cluster of locations), "all subsets" (Neill 2012) (LTSS implemented without proximity or connectivity constraints, returns the highest-scoring unconstrained subset of locations), "ULS" (Patil and Taillie 2004) (returns a high-scoring connected subset based on the Upper Level Set scan statistic within a neighborhood size of k), and "GraphScan" (Speakman et al. 2012) (returns the highest-scoring connected subset within a neighborhood size of k). The expectation-based Poisson (EBP) scoring function (Neill et al. 2005) was used for each of these methods.

Various types of spatial injects were created and randomly inserted in the 2-year time frame of the Emergency Department data. Each of these injects had a duration of 14 days with linearly increasing severity. At a fixed false-positive rate of 1 per month, we recorded the proportion of outbreaks detected and average number of days required to detect an outbreak for each method.

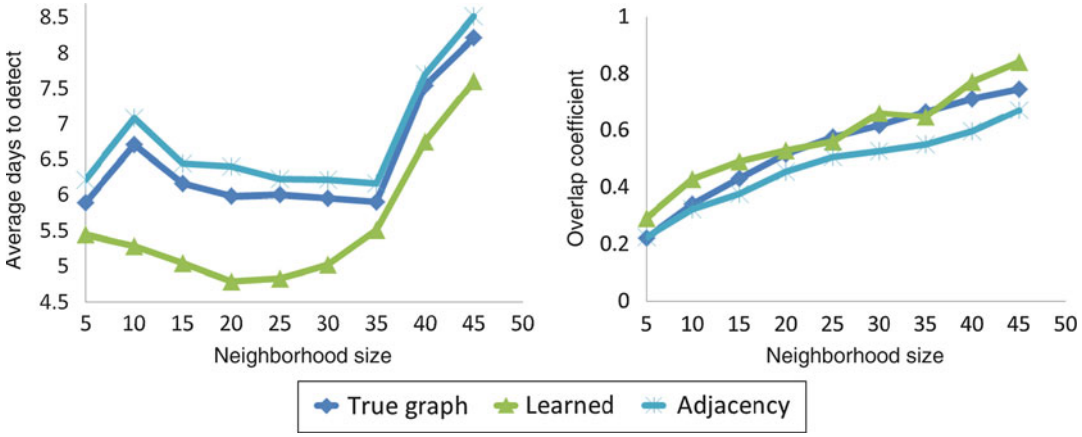
Figure 1 provides the time to detect and overall detection rate for these injects. Averaging across all ten inject types (2,000 outbreaks in total), GraphScan with a neighborhood size of $k = 15$ provides shortest detection time and greatest detection power. ULS provides similar, but slightly lower, detection power due to the fact that it is not guaranteed to identify the highest-scoring connected subset. GraphScan detected 1.79 days earlier than the circular scan and had fewer than half as many missed outbreaks. Note the low detection power of the unconstrained LTSS method. This emphasizes the importance of incorporating spatial and/or connectivity constraints for adequate performance on the spatial event detection task.

Results on Graph Structure Learning

We generated simulated disease outbreaks that spread based on the zip code adjacency graph with additional edges added to simulate travel patterns and injected these outbreaks into our real-world hospital data. We evaluated detection



Disease Surveillance, Case Study, Fig. 1 Average time to detect and overall detection rate for simulated disease outbreaks



Disease Surveillance, Case Study, Fig. 2 Comparison of detection performance of the true, learned, and adjacency graphs, from Somanchi and Neill (2012)

time and spatial accuracy using the learned graphs for these simulated injects (Fig. 2).

This figure also shows the detection performance given the true (adjacency plus travel) graph and the adjacency graph without travel patterns. We observe that the learned graph achieves comparable spatial accuracy to the true graph, while the adjacency graph has lower accuracy. Additionally, the learned graph is able to detect outbreaks over a day earlier than the true graph and 1.5 days earlier than the adjacency graph without travel patterns. Thus, our method can successfully learn the additional edges due to travel patterns, substantially improving detection performance.

Key Research Findings and Conclusions

GraphScan is a novel method for efficient pattern detection that incorporates linear-time subset scanning with connectivity constraints (Speakman et al. 2012). Although similar to the previously proposed FlexScan algorithm, GraphScan is able to scale to much larger graphs of over 100 nodes, with a 450,000-fold increase in speed compared to FlexScan for neighborhoods of size $k = 30$. Along with the enormous speed improvements, GraphScan is guaranteed to identify the highest-scoring connected subset.

The GraphScan algorithm was evaluated against the circular space-time scan statistic (Kulldorff 1997) and the Upper Level Set scan statistic (Patil and Taillie 2004) on synthetic disease outbreaks injected into real-world Emergency Department data from 97 zip codes in Allegheny County, PA. Compared to the competing methods, GraphScan had higher detection power with shorter time required to detect the events, as well as fewer missed events overall.

We also proposed a novel framework which uses GraphScan to learn graph structure from unlabeled data (Somanchi and Neill 2012). This approach can accurately learn a graph structure which can then be used by graph-based event detection methods, enabling more timely and accurate detection of outbreaks which spread based on that latent structure. Our results show that the learned graph structure is similar to the true underlying graph structure. Interestingly, the resulting graph often has higher detection power than the true graph, enabling more timely detection of outbreaks, while achieving similar spatial accuracy to the true graph. This is because the learning procedure is designed to capture not only the underlying graph structure but the characteristics of the events which spread over that graph. By finding graphs where the highest connected subgraph score is consistently close to the highest unconstrained subset score when an event is occurring, we identify a graph structure which is optimized for event detection.

Cross-References

- ▸ Data Mining
- ▸ Social Networks in Emergency Response
- ▸ Social Networks in Healthcare, Case Study
- ▸ Spatio-Temporal Outlier and Anomaly Detection

References

Akoglu L, McGlohon M, Faloutsos C (2010) Oddball: spotting anomalies in weighted graphs. In: PAKDD (2), Hyderabad, pp 410–421

Chakrabarti D (2004) Autopart: parameter-free graph partitioning and outlier detection. In: PKDD, Pisa, Italy, pp 112–124

Kulldorff M (1997) A spatial scan statistic. Commun Stat 26(6):1481–1496

Neill DB (2012) Fast subset scan for spatial pattern detection. J R Stat Soc B 74(2):337–360

Neill DB, Moore AW, Sabhnani MR, Daniel K (2005) Detection of emerging space-time clusters. In: Proceedings of the 11th ACM SIGKDD international conference on knowledge discovery and data mining, Chicago

Patil GP, Taillie C (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environ Ecol Stat 11:183–197

Somanchi S, Neill DB (2011) Fast graph structure learning from unlabeled data for outbreak detection. Emerg Health Threats J 4:11017

Somanchi S, Neill DB (2012) Fast graph structure learning from unlabeled data for event detection (Submitted)

Speakman S, Neill DB (2010) Fast graph scan for scalable detection of arbitrary connected clusters. In: Proceedings of the 2009 international society for disease surveillance annual conference, Miami Beach

Speakman S, McFowland E, Neill DB (2012) Scalable detection of anomalous patterns with connectivity constraints (Submitted)

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geogr 4:11