Research and Educational Activities

In the subset scan framework, our primary goal is to find the subsets of the data which are most anomalous (or that best match some known and relevant pattern) by maximizing the score function F(S). Since an exhaustive search over subsets is computationally infeasible, typical spatial scan methods either restrict the search space, e.g. by searching over circular or rectangular regions, or perform a heuristic search. The former approach has low detection power for regions outside the search space (e.g. elongated or irregular clusters), while the latter does not guarantee that an optimal or near-optimal region will be found. However, we have recently discovered that many pattern detection methods satisfy a property (**linear-time subset scanning**, or LTSS) which allows efficient optimization over all subsets of the data: the highest-scoring (most anomalous or most relevant) of all the exponentially many subsets of the data can be found in linear time, by sorting the data records according to some function and searching only over regions containing the *k* highest-scoring records (letting *k* vary from 1 to the total number of records *N*). This approach enables us to optimize F(S) by evaluating only N of the 2^N possible subsets. More details on research activities are provided in Sections I-X below, applications are described in Sections XI-XIII, and educational activities are described in Sections XIV-XVI.

I. Theory of linear-time subset scanning

Our first paper on linear-time subset scanning (Neill, 2012) has recently been published in the *Journal of the Royal Statistical Society Part B (Statistical Methodology)*, the #1 statistics journal as ranked by impact factor. This paper includes theoretical results on LTSS, the univariate fast subset scan framework, and incorporation of hard constraints on spatial proximity, as discussed below.

In Neill (2012), we propose a general framework for computationally efficient pattern detection based on the linear-time subset scanning (LTSS) property, enabling rapid computation of the highest-scoring subset of records in massive datasets. Let $D = \{R_1 \ldots R_N\}$ be a set of N data records. Let F(S) be a "score function" mapping a subset of data records $S \subseteq D$ to a real number, and let $G(R_i)$ be a "priority function" mapping a data record $R_i \in D$ to a real number. Let $R_{(j)}$ denote the j^{th} highest priority record. Then F(S) and $G(R_i)$ satisfy the LTSS property if $\max_{S \subseteq D} F(S) = \max_{j=1 \ldots N} F(\{R_{(1)} \ldots R_{(j)}\})$. Additionally, F(S) and $G(R_i)$ satisfy the strong LTSS property if, for all $j \in \{1..N\}$, $\max_{S \subseteq D: |S|=j} F(S) = F(\{R_{(1)} \ldots R_{(j)}\})$. In other words, if F(S) satisfies LTSS, the highest-scoring subset of records can be found by sorting the records by priority and then evaluating only N of the 2^N subsets.

We have developed two methods (inclusion and substitution) to prove that functions satisfy the LTSS property. Proof by substitution assumes that, if record $R_i \notin S$ has higher priority than record $R_j \in S$, then $F(S + \{R_i\} - \{R_j\}) \geq F(S)$, i.e. the score of region S will not be decreased by substituting record R_i for R_j. Proof by inclusion assumes that, if record $R_i \notin S$ has higher priority than record $R_j \in S$, then $\max(F(S + \{R_i\}), F(S - \{R_j\})) \geq F(S)$, i.e. if the lower priority record is included in region S then the higher priority record must be included as well. Both proof methods allow us to show that the score function F(S) satisfies the LTSS property, and proof by substitution also allows us to show that F(S) satisfies strong LTSS. We have used these proof methods to show that a large set of commonly used score functions (including Kulldorff's spatial scan statistic and many recent variants and extensions) satisfy the LTSS property, and we have obtained several sets of sufficient conditions for LTSS, as described in Neill (2012) and summarized in the “Results and Findings” section.

We are in the process of investigating many ways in which LTSS will enable computationally efficient event detection, removing some of the major computational barriers faced by subset scanning methods. For example, we were able to find the most anomalous subset of 97 Allegheny County zip codes for a disease surveillance task (monitoring hospital Emergency Department visits with respiratory chief complaints) in approximately 40 milliseconds using LTSS, while a naïve approach (exhaustive search over all 2^{97} subsets of locations) would require over 10^{20} years.

II. Univariate linear-time subset scanning for spatial datasets

Since LTSS only guarantees a solution to the unconstrained (all-subsets) optimization problem, the biggest challenge is to incorporate constraints such as spatial proximity, graph connectedness, or temporal consistency to ensure that relevant and useful subsets are detected. In recent work, we have developed a number of novel and powerful methods for *constrained* optimization, using the unconstrained LTSS method as a building block.

In Neill (2012), in addition to presenting the basic theoretical framework for LTSS described above, we focus on the application of LTSS to univariate spatial data, and consider how spatial proximity constraints can be incorporated. We often want to use spatial information to constrain our search by penalizing or excluding unlikely subsets (e.g. spatially dispersed or highly irregular regions). Thus we propose “fast localized scan” approaches which incorporate spatial proximity constraints into the LTSS framework. For example, we can constrain our search to regions consisting of a center location and some subset of its k-nearest neighbors, using LTSS to reduce the complexity from exponential to linear in k. While these approaches place a hard constraint on the neighborhood size, we also propose “fast localized multiscan” approaches, which do not constrain the neighborhood size but instead penalize larger neighborhoods. As discussed in the “Results and Findings” section below, these efficient, spatially-constrained LTSS searches allow us to perform spatial detection tasks in milliseconds that would require years for exhaustive search, and substantially improve detection power and spatial accuracy as compared to the traditional spatial scan approach (searching over circular regions).

In the past year, we have developed a new framework which allows us to incorporate “soft” constraints on spatial proximity, rewarding compact regions and penalizing sparse regions (Speakman, McFowland, Somanchi, and Neill, 2012). Our previous fast subset scan approach can incorporate proximity constraints using a fixed neighborhood size k; however, each of the 2^k subsets are considered equally likely, and thus the fast localized scan does not take into account the spatial attributes of a subset. Thus we extended the fast localized scan by giving preference to spatially compact clusters while still considering all subsets within a given neighborhood. For a given local neighborhood with center location s_c and size k, we place a bonus or penalty $\Delta_i = h(1 - 2d_i/r)$ on each location s_i, where d_i is that location's distance from the center, r is the neighborhood radius, and h is a constant representing the strength of the compactness constraint. Each Δ_i can be interpreted as the prior log-odds that s_i will be affected, and thus the center location is e^h times as likely as its (k-1)th nearest neighbor. In work presented at the 2011 International Society for Disease Surveillance Annual conference (abstract published in the *Emerging Health Threats Journal*), we demonstrated that this approach can efficiently and accurately detect irregularly-shaped outbreaks. This work will also be presented at the International Workshop on Applied Probability (IWAP 2012) this summer.

III. Linear-time subset scanning for multivariate and tensor datasets

While Neill (2012) focuses on univariate event detection (in which we monitor a single spatio-temporal data stream), linear-time subset scanning can also be extended to multivariate event detection (in which we integrate information from multiple spatio-temporal data streams). In Neill, McFowland, and Zheng (2012, submitted for publication), we demonstrate how LTSS can be used to speed up two different multivariate scan statistic methods, Subset Aggregation (an extension of the method proposed by Burkom et al., 2005) and the multivariate spatial scan proposed by Kulldorff et al. (2007). This work is currently in the 2nd round of review by the journal, *Statistics in Medicine*, and was presented as part of our invited talk on multivariate surveillance at the *2011 Joint Statistical Meetings*.

Subset Aggregation sums the observed and expected counts across the multiple data streams and computes the univariate scan statistic F(S) using these aggregate counts; we have extended this method to search over subsets of data streams, but a naïve implementation would be exponential in the number of streams. In the spatial setting, we also want to search over subsets of locations, so the naïve implementation would be exponential in the number of locations as well. We have developed three computationally efficient solutions, the "fast-naïve" (FN), "naïve-fast" (NF), and "fast-fast" (FF) methods. The key insight behind all three methods is that LTSS can either be used to efficiently optimize a score function over subsets of data streams for a given subset of locations, or to efficiently optimize over locations for a given subset of attributes. The FN method does the former for each possible subset of streams, thus producing a method which is linear in the number of locations but exponential in the number of streams; the NF method does the latter for each possible spatial region under consideration, thus producing a method which is linear in the number of subsets considered (and thus, possibly exponential in the number of locations) but linear in the number of streams. Finally, the FF method iterates between optimizing over subsets of locations and subsets of streams until the algorithm converges to a (local) maximum of the score function over all subsets of locations and streams; multiple restarts are used to approach the global maximum.

Kulldorff's method, instead of aggregating counts across the multiple data streams, computes the univariate score F(S) separately for each data stream and then sums these log-likelihood ratios across the multiple streams, thus making the implicit assumption that the data streams are conditionally independent given the affected region S. A naïve implementation of Kulldorff's method is linear in the number of data streams, but is also linear in the number of subsets considered (and thus, possibly exponential in the number of locations). Thus we have developed a fast implementation of Kulldorff's method (FK) which is linear both in the number of streams and number of locations. To do so, we use the fact that the parametric statistics satisfy the LTSS property (and thus, can be efficiently optimized over subsets of locations) for any fixed values of the relative risk parameters for each data stream. Thus we can iterate between optimizing over subsets of locations given the fixed relative risks, and optimizing the relative risks for the given subset of locations. As in the FF method, this procedure converges to a local maximum of the score function, and multiple restarts are used to approach the global maximum.

In Neill, McFowland, and Zheng (2012), we evaluate the run time of the fast and naïve versions of each method, demonstrating that our LTSS-enabled algorithms make multivariate subset scanning computationally feasible even when the numbers of locations and data streams are large. We also perform a detailed comparison of the detection power and spatial accuracy of the Subset Aggregation and Kulldorff methods, using our efficient algorithms for each method. These results are described in the "Results and Findings" section.

We have recently extended the multivariate LTSS approach from matrix data (records x attributes) to **tensor data** with an arbitrary number of modes (Kumar and Neill, 2012, submitted for publication). Our previous work allows efficient optimization over subsets of records and attributes, which can be thought of as the rows and columns of a matrix; the current work allows joint optimization over subsets of each mode of a tensor with three or more modes. Our approach is a natural generalization of our multivariate fast subset scan algorithm: we randomly initialize the algorithm, then iteratively optimize over subsets of each tensor mode given the other modes. Each such conditional optimization can be performed efficiently using the LTSS property; the iterative process converges to a local maximum of the score function, and then multiple randomized restarts can be used to approach the global maximum. In a recent paper submitted to the 2012 IEEE International Conference on Data Mining, we applied this technique to the disease surveillance domain, using multivariate case data from individuals in a population. In this setting, we have not only multivariate spatio-temporal count information (the number of cases for each location, time step, and data stream), but also additional categorical attributes for each affected individual (such as age group and gender). Each such attribute is represented by a tensor mode, and location and data stream are represented by two additional modes. Our Fast Tensor Scan (FTS) approach identifies not only the affected spatial locations and data streams, but also the characteristics of the affected subpopulation, as represented by a subset of values for each monitored attribute (e.g. "males under 30 who use intravenous drugs"). We demonstrate that this approach enables accurate and timely detection of emerging events, while maintaining computational tractability for massive datasets, as discussed in the "Results and Findings" section below.

Our ongoing work involves applying FTS to many other application domains, including container shipment monitoring and detection of anomalous patterns of care in a hospital setting. One challenge in non-spatiotemporal domains is estimating the expected count corresponding to each combination of attributes in the tensor; our ongoing work learns the structure and parameters of a Bayesian network from the case data and uses the learned model to predict counts. We believe that the ability of FTS to accurately characterize the areas and subpopulations affected by an event will have important applications in other areas, such as disaster and crisis response, where it is essential for an intervention to be both rapid and precisely targeted. We are also working with the Chicago Department of Public Health to obtain data about the prevalence of sexually transmitted illnesses, risk factors, treatments, and preventive measures, in order to detect emerging patterns of STIs which may differentially affect a particular subpopulation or a particular neighborhood.

IV. GraphScan: linear-time subset scanning with graph connectivity constraints

Another extension of linear-time subset scanning focuses on graph and network data, where we monitor one or more data streams at each node of the graph, and wish to detect the most anomalous subset of nodes subject to the graph connectivity constraints (i.e. the given subset of nodes must form a connected subgraph of the original graph). If the score function satisfies LTSS, we can prove the following rule: if a group of nodes is contained in the optimal subset S, and if removing that group does not disconnect the subgraph, any neighboring group of nodes with higher priority will be contained in S as well. In the past year, we have developed a new and more computationally efficient version of the GraphScan algorithm which relies on a depth-first backtracking search rather than evaluating nodes in priority order (Speakman, McFowland, and Neill, 2012, submitted for publication). This latest version is able to scale to graphs of several hundred nodes, and we have demonstrated that both detection power and spatial accuracy are substantially improved when incorporating both connectivity and proximity constraints as compared to proximity constraints alone. GraphScan can be used for spatial data (searching for the most anomalous connected cluster of zip codes, with edges defined by spatial adjacency, travel patterns, etc.), and can also be used for non-spatial data with an underlying graph structure (including cell phone call graphs, social networks, and the Enron e-mail dataset). We believe that this approach will be particularly useful for our future work on detecting and preventing hospital-acquired illness, monitoring the spread of nosocomial infections between hospitals and between rooms within a hospital based on the movement of patients and hospital staff.

This summer, our GraphScan algorithm will be presented at the Quality and Productivity Research Conference (QPRC 2012), and a related book chapter is under review for the *Encyclopedia of Social Network Analysis and Mining*. The full paper is currently under revision, and will be re-submitted to the journal, *IEEE Transactions on Knowledge and Data Engineering* (TKDE). The GraphScan optimization step was also embedded into our tensor scan approach described above, thus allowing us to place a connectivity constraint (with a given graph structure) on our search over subsets for any or all modes of the tensor. GraphScan has also been extended to incorporate temporal consistency constraints and to learn the underlying graph structure, as described below.

V. Using GraphScan to learn the underlying graph structure

While graph-based event detection algorithms (such as our GraphScan approach described above) typically assume a known graph structure, we have developed a new, general framework for **learning an unknown graph structure** from data, and demonstrated that the graph learning framework enables more timely and more accurate event detection (Somanchi and Neill, 2012, submitted for publication).

Processes such as disease propagation or information diffusion often spread over some latent network structure (e.g. social networks or person-to-person contacts) which must be learned from our observations of the nodes in the network. For example, in disease surveillance, we might observe the time series of case counts for each of a set of spatial locations, but not know which locations are likely to spread disease to which other locations (via spatial adjacency, travel patterns, common food or water sources, etc.). Thus we attempt to reconstruct the underlying network along which a disease outbreak or other event might spread, and use the learned network to improve the timeliness and accuracy of event detection. However, in many cases labeled data may not be available: for example, public health officials might be aware that an outbreak has occurred, but may not have detailed information about which areas were affected and when. Hence we focus on learning graph structure from **unlabeled** data, given only a time series of observed counts (such as hospital visits or medication sales) at each node.

Our solution is to compare the most anomalous subsets detected with and without the graph constraints: we score each of a set of potential graph structures $G_1 \ldots G_M$ for each training example $D_1 \ldots D_J$, finding the most anomalous (highest scoring) connected subset and its score using an efficient graph-based event detection algorithm (our GraphScan algorithm, or the fast but approximate ULS approach) for each combination of graph structure and dataset. These scores are normalized by dividing by the score of the most anomalous unconstrained subset for that training example, which can be efficiently computed using LTSS, and the normalized scores for a potential graph structure G_m are averaged over all of the J training examples. The idea is that, if the given graph structure is close to the true underlying graph structure, then the maximum constrained score will be close to the maximum unconstrained score for many of the training examples, while if the graph structure is missing essential connections, then the maximum constrained score given that graph structure will be much lower than the maximum unconstrained score for many examples. However, any graph with a very large number of edges will also score very close to the maximum unconstrained score, and thus we compare the score of the given graph structure to the distribution of scores of random graphs with the same number of edges, and choose the graph structure with the most statistically significant score given this score distribution.

Within our general framework for graph structure learning, we compared five approaches which differed both in the underlying detection method (BestSubgraph) and the method used to choose the next edge for removal (BestEdge), incorporated into a provably efficient greedy search procedure. We demonstrated both theoretically and empirically that our framework requires fewer calls to BestSubgraph than a naive greedy approach. We also evaluated the scalability, detection power, and accuracy of these approaches on various types of simulated disease outbreaks, including outbreaks which spread according to spatial adjacency, adjacency plus simulated travel patterns, and random graphs (Erdos-Renyi and preferential attachment), as discussed in the "Results and Findings" section.

This work has been completed and will be submitted to the IEEE International Conference on Data Mining next month. It was also presented at the 2011 International Society for Disease Surveillance Annual Conference (abstract published in the *Emerging Health Threats Journal*). Our ongoing work focuses on extending the graph structure learning framework in several directions, including learning graph structures with directed rather than undirected edges, learning graphs with weighted edges, and learning dynamic graphs where the edge structure can change over time.

VI. Detecting dynamic events with temporal consistency constraints

In recent work (Speakman and Neill, in preparation), we have developed a novel method for incorporating **soft constraints** into our linear-time subset scanning framework. Unlike many of the LTSS approaches describe above, we do not restrict the search space, but instead consider all subsets of the data while rewarding subsets that are more likely or penalizing subsets that are less likely to be affected. Incorporating soft constraints into the LTSS framework is challenging because, for an arbitrary score function F(S) that satisfies the linear-time subset scanning property, a penalized version of that function is not guaranteed to satisfy LTSS.

However, we have shown that this problem can be circumvented by conditioning on the event's severity (or relative risk), denoted as *q*. For a given value of *q*, and assuming any expectation-based scan statistic in the separable exponential family, we have shown that the score function $F(S \mid q)$ can be written as an additive function, $F(S \mid q) = \sum_{s_i \in S} G_q(s_i)$. For such functions satisfying the **additive LTSS property**, we can write the penalized form $F_{pen}(S \mid q) = \sum_{s_i \in S} (G_q(s_i) + \Delta_i) = \sum_{s_i \in S} H_q(s_i)$, where $H_q(s_i)$ is the total contribution of location s_i to the penalized scan statistic for the given value of *q*. Thus, for a given severity value *q*, we can easily maximize $F_{pen}(S \mid q)$ over all subsets of the data, by choosing all and only those locations with positive values of $H_q(s_i)$. Connectivity constraints can also be incorporated into this framework: this becomes an NP-hard problem (minimum Steiner tree), but is still feasible for graphs with several hundred nodes.

By applying the additive LTSS property, we can enforce soft constraints on **temporal consistency** by considering the patterns detected at adjacent time steps, and rewarding patterns that are not dramatically different between time steps t and t+1. This allows us to extend our detection methods from detecting static patterns (which affect a fixed set of locations for some time duration) to **dynamic patterns** (which can grow or spread over time) while still maintaining efficient computation. Additionally, by using temporal consistency constraints to share information between multiple time steps, we can allow patterns to evolve smoothly over time while penalizing patterns which display unrealistic temporal trends (e.g. affecting the east side of the city on day 1, the west side on day 2, and back to the east side on day 3).

A preliminary version of this approach, presented at the 2011 INFORMS Annual Conference, applies the temporal consistency constraints moving forward in time, rewarding locations which were present for each of the past two time steps and also the neighbors of these locations. This algorithm was used to detect dynamic patterns in graph data with connectivity and temporal consistency constraints, applied to the detection of **spreading contaminants in a water distribution network**. Our preliminary results show that incorporating simple size and temporal consistency constraints in a penalized, expectation-based binomial scoring function allows GraphScan to detect the contaminants earlier and to more accurately identify which nodes are affected as the contamination spreads through the network.

Our current extensions of the algorithm, which will be incorporated into a journal paper submission in the next few months, are threefold. First, instead of only propagating our beliefs about the affected subset forward through time, we have developed and are currently evaluating an iterative approach which enables propagation of information both backward and forward in time. Second, we have developed new methods for choosing the deltas (bonuses or penalties) for including each location on a given time step, based on a generative model of event propagation. Finally, we have developed new approaches for optimizing over the space of possible values of the severity parameter *q*. We believe that these extensions will dramatically improve detection performance for emerging dynamic events which grow, shrink, or move over time.

VII. FGSS: linear-time subset scanning for general multivariate datasets

In McFowland, Speakman, and Neill (2012, submitted for publication), we present Fast Generalized Subset Scan (FGSS), an extension of the LTSS approach which enables efficient pattern detection in general multivariate datasets. In this case, we do not have space-time data, but instead have an arbitrary set of attributes measured for each of a large set of data records. In this problem setting, our goal is to detect self-similar subsets of data records for which some subset of attributes are anomalous. Our approach consists of four steps: 1) efficiently learning a Bayesian network which represents the assumed null distribution of the data; 2) computing the conditional probability of each attribute value in the dataset given the Bayes Net, conditioned on the other attribute values for that record; 3) computing an empirical p-value range corresponding to each attribute value by ranking the conditional probabilities, where under the null hypothesis we expect empirical p-values to be uniformly distributed on [0,1]; and 4) using a nonparametric scan statistic to find subsets of records and attributes with an unexpectedly large number of low (significant) empirical p-values. The final step is computationally expensive (exponential in the numbers of records and attributes for a naïve search), but LTSS can be used to speed up this search as in the spatial setting discussed above, converging to a local maximum of the score function and ensuring that each iteration step is linear (not exponential) in the number of records or attributes. The FGSS framework has been evaluated on several application domains, including early detection of simulated anthrax bio-attacks, discovery of patterns of illicit container shipments for customs monitoring, and network intrusion detection, demonstrating improved detection accuracy, efficient runtime, and ability to correctly characterize the affected subset of attributes in all three domains. Results of these evaluations are presented below.

In the past year, we have completed our first paper on FGSS (submitted to the *Journal of Machine Learning Research*, and currently under review), as well as presenting talks at the *International Workshop on Applied Probability (IWAP 2012)* and the *2011 INFORMS Annual Conference*. We have also extended the FGSS approach to the case of **mixed real- and categorical-valued datasets**, augmenting the Bayesian network with a regression tree for each real-valued attribute. Then the probability density corresponding to each attribute value (conditioned on the other attribute values for that data record) is computed by performing kernel density estimation using only the appropriate leaf of the regression tree. We can then compute the empirical p-values for that attribute by computing and ranking the kernel density estimates corresponding to each attribute value, and perform the fast LTSS-enabled nonparametric scan as before. We are also investigating an alternative model which uses a **dependency network** instead of a Bayesian network to model the distribution of the data when no events are occurring; this avoids the computationally expensive Bayes Net structure learning step, and enables the algorithm to scale to datasets with hundreds or thousands of attributes, making it usable for our work on discovering anomalous patterns of patient care discussed below.

VIII. Semantic scan statistic

Typical event detection systems aggregate data records into counts and then detect spatial areas with anomalous recent counts. For example, in disease surveillance, we count the number of disease cases with each of a small set of general symptom categories (such as respiratory, gastrointestinal, and fever) in each zip code for each day. This approach works reasonably well given limited data about each patient, but we believe that outbreak detection can be dramatically improved by incorporating rich text data from electronic health records, e.g. patient histories and chief complaints. Typical disease surveillance systems have difficulty detecting new emerging infections with unknown symptom patterns, or other diseases that do not correspond to the existing symptom categories. Thus we are exploring a new "semantic scan statistic" approach, which uses rich text data to detect previously unknown event types, forming and searching a huge number of aggregated count datasets on the fly (Liu and Neill, 2011; Murray, Liu, and Neill, in preparation). Each count represents the number of records in a given spatial area and time interval which match some set of keywords; different keyword sets are used for each aggregation. Since the number of possible keywords is huge, our challenge is to find the most interesting aggregations and anomalous subsets without an exhaustive search. Our current approach uses topic models (created by Latent Dirichlet Allocation) to automatically discover possibly relevant subsets of keywords. We then form a separate count dataset from the case data for each topic, and find the maximum region score over all of the topics considered. Thus the semantic scan statistic provides information not only about whether an event has occurred and which space-time region has been affected, but also which set of keywords (topic) occurs with surprisingly high count in this region. Our goal is to show that, for disease outbreaks with very specific sets of symptoms, or with novel combinations of symptoms that have not previously been seen in the data, the text-based analysis will enable earlier and more accurate detection than traditional count-based detection approaches.

Our approach assumes that there is a latent "topic" in each case report, and thus we can utilize widely used topic models to extract those "topics" from text and then apply existing spatial scan techniques to them. For example, we might have one patient with "abdominal pain and nausea" symptoms, and another patient who has exhibited "vomiting", but both sets of symptoms might correspond to the same disease category (GI illness). Our first approach used topic models to extract some number of "static" topics from the entire training dataset, and additional "dynamic" topics learned from the current two weeks of test data, in order to capture both broad, typical syndrome categories and newly emerging trends in the recent data. For each day of data, for each of the extracted topics, we formed a count dataset from the case data by computing the number of cases in each zip code for each day which are most likely to correspond to the given topic. We then apply our novel spatial scan methods to the resulting count data, and report the maximum value of the spatial scan statistic over all topics.

Using a combination of "static" and "dynamic" topics does reasonably well for picking up patterns of symptoms corresponding to both typical and (simulated) newly emerging illnesses, but some detection power is lost because many of the dynamic topics do not capture sufficiently different syndrome groupings from those represented by the static topics. Thus our current approach focuses on learning "incremental" topics that represent those trends in the current data which are not well captured by the “static” topics, and we demonstrate that the resulting incremental Latent Dirichlet Allocation approach shows substantially improved detection power for newly emerging illnesses.

Our preliminary work was presented at the 2011 International Society for Disease Surveillance Annual Conference (abstract published in the *Emerging Health Threats Journal*) and at the 2012 International Conference on Digital Disease Detection. We anticipate submitting the full paper to the IEEE International Conference on Data Mining (ICDM 2012) this summer. We are also working to further improve detection power by incorporating spatial information into the topic modeling step (rather than just into the subsequent spatial scan), and are performing an in-depth evaluation and comparison of the different variants of the method. Future work will also apply the semantic scan method to informal, online data sources such as Twitter and evaluate the utility of such sources for disease surveillance.

IX. Bayesian fast subset sums

The multivariate Bayesian scan statistic (MBSS) is a powerful detection method which can integrate information from multiple data streams and can model and distinguish between multiple event types (Neill and Cooper, 2010). The output of the MBSS method can be easily visualized by computing the posterior probabilities that each event type E_k has affected each spatial location s_i, summing the posterior probabilities for all regions S containing s_i. Unlike standard spatial scan visualizations, which do not compute probabilities but instead show the most likely cluster, this method is able to quantify the system’s uncertainty about the spatial extent and type of events. However, our LTSS method cannot be used to efficiently generate this visualization, since we need to sum over probabilities rather than just finding the highest-scoring region.

Thus we have developed an efficient “Fast Subset Sums” (FSS) method which computes the summed posterior probability over all subsets containing location s_i, without computing the posterior probability of each individual subset (Neill, 2011). This work extends the MBSS framework to enable detection and visualization of irregularly-shaped clusters in multivariate data, by defining a hierarchical prior over all subsets of locations. While a naive search over the exponentially many subsets would be computationally infeasible, we demonstrate that the total posterior probability that each location has been affected can be efficiently computed, enabling rapid detection and visualization of irregular clusters. We compare the run time and detection power of this “fast subset sums” (FSS) method to our original MBSS approach (assuming a uniform prior over circular regions) on semi-synthetic outbreaks injected into real-world Emergency Department data from Allegheny County, PA. We demonstrate substantial improvements in spatial accuracy and timeliness of detection, while maintaining the scalability and fast run time of the original MBSS method. Results are summarized in the “Results and Findings” section, and the full paper was recently published in the journal *Statistics in Medicine* (2011).

We have recently developed a generalization of the fast subset sums method which allows the sparsity of the detected region to be controlled (Shao, Liu, and Neill, 2011). More precisely, we propose a hierarchical probabilistic model with three steps: first, choosing the center location s_c from a multinomial distribution; second, choosing the neighborhood size k from a multinomial distribution; and third, independently choosing whether to include (with probability p) or exclude (with probability 1-p) each location in the k-neighborhood of the center. We demonstrate that our previously proposed MBSS and FSS methods correspond to special cases of this **Generalized Fast Subset Sums** (GFSS) method, with p = 1 and p = 0.5 respectively, and show that appropriate choice of the sparsity parameter p enables much faster detection and higher spatial accuracy than either MBSS or FSS. Finally, we demonstrate that the distribution of the sparsity parameter can be accurately **learned** from a small amount of labeled training data, and that the resulting GFSS method with learned p distribution outperforms MBSS, FSS, and GFSS with a uniform p distribution. We also show that two otherwise identical event types with different sparsities can be reliably distinguished by learning each event's p distribution, and that learning both an event's sparsity distribution and its relative effects on different data streams leads to more timely detection and better characterization than learning either parameter on its own.

In the past year, this work was presented at, and published in the proceedings of the *IEEE International Conference on Data Mining* (ICDM 2011). Future extensions of this work will jointly learn the distributions for the sparsity parameter, neighborhood size, and center location using expectation-maximization (EM). We are also working on extensions to partially labeled data, where only a small subset of the affected locations is provided.

X. Maximizing cross-correlation over subsets of locations and streams

We have developed a novel method which extends our linear-time subset scanning framework to **identify spatially localized subsets of leading indicators** for event prediction (Flaxman and Neill, 2012, submitted for publication). Given a spatially localized time series to be predicted (e.g. daily counts of violent crime for each census tract) and multiple potential predictors (e.g. daily counts of various types of calls for service for each tract), our approach maximizes the cross-correlation between the predictor variable and an aggregated subset of leading indicators across a range of time lags, all subsets of potential predictors, and all proximity-constrained subsets of locations. This captures the fact that different subsets of leading indicators may be relevant in different areas of the city. However, even for relatively small numbers of locations and leading indicators, optimization over all such subsets is computationally infeasible, and unfortunately the function we wish to optimize (Pearson correlation between the independent and dependent time series) does not satisfy our linear-time subset scanning property. Instead, we propose a novel "iterative average dot product" (IADP) method, with the key insight that both spatial subset search and feature selection can be performed by approximating the correlation with a function that can be efficiently maximized over subsets of locations or streams using LTSS. We then iterate between conditionally optimizing the approximate correlation over subsets of locations (for a given subset of streams) and optimizing over subsets of streams (for a given subset of locations), until convergence to a joint local optimum. Our iterative procedure refines the quality of this approximation over time, approaching the true best correlation. The approach was tested on 311 service calls from Chicago, and compared both to ground truth (for small problem sizes) and existing feature selection methods (such as LASSO regression). Our method found near-optimal correlations while scaling to large numbers of locations and data streams, and demonstrated significant improvements in the correlation of detected subsets as compared to existing methods. This work will be presented at the International Symposium on Forecasting this summer, and the full paper has been submitted to the IEEE International Conference on Data Mining (ICDM 2012).

XI. Application domain #1: Early detection of disease outbreaks

The **disease surveillance** domain has served as our primary testbed for the development of new event detection methods; we have obtained real public health data from a number of sources, and our work has been incorporated into multiple deployed biosurveillance systems. We are currently collaborating with a team led by Dr. Rick Davies of the Ottawa Heart Institute, developing and deploying systems for real-time disease surveillance in the Grey Bruce region of Ontario (ECADS project), Ottawa (ASSET project), and several other Canadian cities (Data Fusion project). The ECADS and ASSET systems are essentially complete, and the deployed systems are in regular use by Grey Bruce Public Health and Ottawa Public Health respectively. The ECADS system has enabled Grey Bruce to rapidly detect several emerging outbreaks (*E. coli*, scarlet fever, cryptosporidiosis, etc.) and our work has given both health departments powerful tools for outbreak investigation and response. The Data Fusion project will focus on development and deployment of methods for fusion of multiple health data sources, with applications to the monitoring of **hospital-acquired infections** and detection of patterns of events related to **drug abuse**. Additional biosurveillance system deployments (in collaboration with CMU's Auton Lab) are in progress in Sri Lanka and in Tamil Nadu, India.

The project PI, Daniel Neill, recently published an article on "New directions in artificial intelligence for public health surveillance" in *IEEE Intelligent Systems*. This article, based in part on the semantic scan statistic and fast subset scan methods described above, focused on the need for methods that can incorporate rich, unstructured text data and can scale to huge numbers of records and data streams.

In addition to continuing to work with our existing Allegheny County Emergency Department data as a testbed for development of new algorithms, and working with the ECADS/ASSET/Data Fusion project team to develop and deploy methods for monitoring of multiple public health data sources in Canada, we are currently working to acquire new data and establish broader collaborations with the Chicago Department of Public Health. As noted above, our first project will involve analysis of sexually transmitted illness data using our new tensor scan methods. We are also working with the Data Fusion team to acquire and analyze hospital-acquired illness data, which also includes detailed information on patient location and movements within the hospital, and we anticipate that our dynamic GraphScan methods will be very useful in this setting.

XII. Application domain #2: Detection and prediction of patterns of crime

In recent work with Prof. Wil Gorr, we have applied the expectation-based scan statistic approach to crime hot-spot detection using offense report data. In addition to successfully detecting spatial clusters of violent crime, we demonstrated that the scan statistic can also be used to accurately predict clusters of violent crime between 1 and 3 weeks in advance, by detecting clusters of less serious "leading indicator" crimes. This early warning has the potential to enable police to reduce crime through reallocation of patrols and other targeted interventions.

In 2009 we began collaborating with the Chicago Police Department, and they have recently provided us with a huge amount of data including crime offense reports, calls for service, web-based citizen complaints, curfew violations, contact cards, gang information, and deployment operations areas (representing targeted interventions by the CPD to prevent crime). We are in the process of analyzing this data to determine which crime types and other indicators are the best predictors of emerging hot-spots of violent crime, both individually and when integrated with other data sources, to evaluate and extend our methods for prediction of violent crime patterns, and to develop and compare other methods for crime prediction.

Moreover, we have developed and provided a software package (which we call "CrimeScan") to the CPD, which uses our multivariate spatial and subset scan methods to monitor multiple leading indicators for crime prediction, and their use of this tool has grown substantially over the past year. Our CrimeScan software is now in day-to-day use by the CPD for crime prediction and tactical deployments. In the past year, our work in this area has proceeded in three directions:

First, we are in the process of performing a detailed **empirical evaluation** of the existing version of CrimeScan and comparison to existing crime prediction methods in the literature, using crime offense report and 911 call data supplied by the CPD (Neill, Nagasunder, Goldstein, Lee, and Gorr, in preparation). Our results suggest that CrimeScan is able to predict 60% of the "clustered" violent crime at a false positive rate of 15%, which improves substantially on previous methods based on kernel density estimation (commonly used in the law enforcement domain). We anticipate that our results will be complete, and a journal paper submitted for publication, this summer.

Second, we are currently **extending the CrimeScan prediction model**, and anticipate that this approach will further improve prediction performance (Wu and Neill, in preparation). Our original approach detects clusters of leading indicators and then uses these to predict that violent crime will occur nearby. Our current approach extends these by learning a logistic regression model to predict the probability that a violent crime cluster will occur, as a function of location and the presence of various types of leading indicator clusters. These results will either be incorporated into the first CrimeScan paper mentioned above, or will be submitted as a second paper in the fall.

Finally, in collaboration with Brett Goldstein (Chief Data Officer of the City of Chicago), we are working to extend our event prediction work in order to predict many other quantities that are relevant to the city. Our CrimeScan software has been extended and re-dubbed "CityScan", and is currently being evaluated for predicting trends in 311 calls for service, trends in Twitter data related to the city's public transportation, and check-ins on FourSquare, a location-based social network. Plans are to have our software running in real time to detect relevant trends and patterns which will be directly used by city services. We are very excited about this ongoing collaboration, which has the potential to make a significant, data-driven contribution to city management in practice.

XIII. Application domain #3: Detecting anomalous patterns of patient care

Dr. Neill was recently awarded a Healthcare Technology Innovation grant from the University of Pittsburgh Medical Center (UPMC) Technology Development Center, with the goal of "Anomalous Pattern Detection in Healthcare Data Streams." As part of this project, we plan to apply novel pattern detection methods to healthcare data flowing through UPMC's Message Routing System, thus enabling the ongoing extraction of diagnostic and business intelligence from real-time healthcare data streams.

While the proposed system is general enough to detect many different types of patterns, our primary focus will be **detecting anomalous patterns of care with significant impacts on patient outcomes**. Consider the natural variation in care practices between different groups and different clinicians within the UPMC systems. For example, when presented with a patient with severe breathing difficulties, different clinicians may choose to administer different types and dosages of medications, use different criteria to decide whether or not to place the patient on a ventilator, etc. Similarly, different hospital staff may exhibit variability in their care practices (such as hand-washing and isolation precautions) and adherence to physician orders. This variation in type and quality of care can have huge impacts on patient outcomes, such as mortality and morbidity rates, hospital re-admissions, and nosocomial infections. The proposed system will evaluate, extend and deploy our state-of-the-art pattern detection methods, in order to automatically detect substantial variations in care between groups which have significant impacts on patient outcomes. These impacts can either be negative (e.g. systematic errors), in which case we can detect and correct these sub-optimal patterns of care, or positive. In the latter case, our system will have discovered a new potential best practice, which can then be investigated further, and if appropriate, shared with other groups.

Our first stage is to analyze existing data from UPMC consisting of thousands of patients with an entering diagnosis of chronic obstructive pulmonary disorder (COPD), and discover anomalous patterns of care for these patients which influence outcomes such as length of stay, readmission, mortality, and nosocomial illness (e.g. ventilator-acquired pneumonia and multidrug-resistant infections). Due to personnel turnover on the UPMC side, as well as delays with the data use agreement between CMU and UPMC, our data acquisition has been delayed, but we finally expect to receive usable data in the next month or so. We have also begun talking with Presence Healthcare (a major health provider in the Chicago area with 12 hospitals, 29 long term care and senior residential facilities, numerous outpatient services and clinics, etc.) about partnering on a similar project to identify and investigate anomalous patterns of care for patients with congestive heart failure.

Other application areas

With former CMU Civil and Environmental Engineering graduate student Daniel Oliveira, and CEE faculty members Lucio Soibelman and Jim Garrett, we applied the novel spatial cluster detection methods described above to **detect and explain clusters of pipe breaks** in a water distribution system. A journal paper on this work was recently published in the *Journal of Computing in Civil Engineering* (Oliveira et al., 2011).

XIV. Educational activities: The Machine Learning and Policy (MLP) Initiative

With the critical importance of addressing global policy problems ranging from disease pandemics to crime and terrorism, and the continuously increasing size and complexity of policy data, the use of machine learning has become increasingly essential for data-driven policy analysis and for development of new, practical information technologies that can be directly applied for the public good. The numerous challenges facing our world will require broad, successful innovations at the intersection of machine learning and public policy. This endeavor will require widespread collaboration between machine learning and policy researchers, increased emphasis on the education of future researchers with in-depth knowledge of both disciplines, and a broadly shared research focus on developing novel machine learning methods which directly address critical policy challenges. Thus this project's main educational focus is on a multi-pronged curricular program, the **Machine Learning and Policy (MLP) initiative**. This program will facilitate the widespread use of machine learning methods for the public good by incorporating machine learning throughout the public policy curriculum.

The components of this program include:

<u>Joint Ph.D. Program in Machine Learning and Public Policy</u>. This new doctoral degree program, jointly administered by MLD and Heinz, had its first entering student in Fall 2011. The primary goal of this program is to train future leaders in MLP. It will provide students with an in-depth understanding of both machine learning and public policy, as well as a shared focus on creating novel machine learning methods and applying them toward the public good. The program will also encourage collaboration between students and researchers in machine learning and policy: MLP students will be co-advised by faculty from MLD and Heinz, and will be active participants in both ML and Heinz Ph.D. programs.

<u>Large Scale Data Analysis for Public Policy</u>. This introductory lecture course, developed and taught by Dr. Neill, is geared toward policy students with little or no prior machine learning experience. Students learn to translate policy problems into the "ML paradigms" of prediction, modeling, and detection, choose and apply the appropriate machine learning methods and software tools, and correctly interpret the results. The course also includes a group project, in which students analyze real-world policy data in a domain of their choice, and multiple "guest mini-lectures" from faculty and Ph.D. student researchers, each describing the successful application of machine learning to a specific policy problem. Some of the research methods developed in this project have been presented as part of the "detection" component of the Large Scale Data Analysis course, and the course provides the basic tools and methods necessary for students to begin working on this project or other research projects in machine learning, data mining, and pattern detection.

<u>Research Seminar in Machine Learning and Policy</u>. This seminar, developed and taught by Dr. Neill, is intended for Ph.D. students in Heinz, MLD, and other departments who wish to engage in cutting-edge research at the intersection of machine learning and policy. Students present and discuss current research articles on a range of MLP topics, address fundamental challenges in MLP research, and explore promising research directions through presentation and discussion of ongoing research projects. Research collaborations between machine learning and policy students are explicitly encouraged; these projects often continue beyond the scope of the course and are integrated into the students' thesis work. Research methods and publications developed in this project have formed the basis of several in-depth discussions as part of the Research Seminar, and several student projects have contributed toward the research goals of this project.

<u>Special Topics in Machine Learning and Policy</u>. This half-semester, Ph.D.-level course series, developed by Dr. Neill, will provide in-depth discussion and exploration of a different, specific MLP topic each year. Students will engage in individual or small-group research projects intended to understand and advance

the state of the art.   In 2010-2012, Dr. Neill taught the first three sections of this course, focusing on the topics of "Event and Pattern Detection", "Machine Learning for the Developing World", and "Harnessing the Wisdom of Crowds" respectively. Research methods and publications developed in this project formed the basis of many of the topics discussed in the Event and Pattern Detection course, and several student projects for these courses have contributed toward the research goals of this project.

In the past year, our first student (Seth Flaxman) joined the Joint Ph.D. program in Machine Learning and Public Policy, and has been extremely productive as a member of the PI's Event and Pattern Detection Laboratory. Additionally, the PI was actively involved in coordinating this year's joint Ph.D. program admissions and in leading a faculty search in "Societal-Scale Data Analysis" which successfully recruited an additional tenure-track faculty member in this area. MLP courses taught by the PI in the past year included Large Scale Data Analysis for Public Policy, the MLP Research Seminar, and a new Special Topics in MLP course focusing on "Harnessing the Wisdom of Crowds".

In the near future, in addition to continuing with course and program development, directing the joint Ph.D. program, and being involved with additional faculty hiring, we intend to create an integrated web resource for Machine Learning and Policy at CMU, and to develop a workshop series which will spread these ideas outside the university. As a first step toward wider dissemination of these ideas, the PI is collaborating with several other CMU faculty members to put together an internal workshop on "Machine Learning for the Social Sciences" to be held in October, with invitees across many CMU departments.

XV. Educational Activities: The Event and Pattern Detection Laboratory

Over the past two years, thanks in large part to the generosity of the National Science Foundation, the PI has founded and directs the **Event and Pattern Detection Laboratory** (EPD Lab) at CMU. The Lab currently consists of Dr. Neill and seven students for whom he is the principal research advisor: three Heinz College Ph.D. students (Skyler Speakman, Edward McFowland III, and Sriram Somanchi), one student in the Joint Ph.D. Program in Machine Learning and Policy (Seth Flaxman), one SCS master's student (Kenton Murray), one MSIT-Very Large Information Systems student (Xin Wu), and one MISM student (Yating Zhang). All of these students are working on research highly relevant to the goals of this project, as described in the sections above, and at least two additional students will join the project in the coming year. Strong emphasis is placed on student mentoring and group cohesion: Dr. Neill meets individually with each student on a weekly basis, along with bi-weekly lab meetings and occasional social events. Five additional master's students involved with this project have successfully completed their studies in the past year, including LTI student Yandong Liu, MSIT-VLIS students Amrut Nagasunder and Tarun Kumar, secondary MS in Machine Learning student Kan Shao, and MISM student Rajas Lonkar. All of these students were very successful in their job search and now have positions in industry or government. Notable accomplishments by EPD Lab students related to this project in the past year include:

- Heinz Ph.D. student Skyler Speakman, for his research on dynamic GraphScan applied to detection of spreading contamination in a water network, was the 2011 runner-up for the George Duncan

Award, given yearly by CMU's Heinz College to the best "Second Research Paper" among the 3rd year students.
- Heinz Ph.D. student Ed McFowland, for his research on Fast Generalized Subset Scan, was the 2012 winner of the Suresh Konda Award, given yearly by CMU's Heinz College to the best "First Research Paper" among the 2nd year students.
- Research related to this project led to the completion of Sriram Somanchi's Heinz First Research Paper, Tarun Kumar's MSIT Capstone Project, Amrut Nagasunder's MSIT Capstone Project, and Kan Shao's Data Analysis Project (equivalent to a master's thesis for the MS in Machine Learning).

While we have presented multiple papers and talks under the EPD Lab banner in the past year, in the next year we hope to give the lab greater visibility by putting together a high-quality website, making more of our software publicly available, and inviting our faculty research collaborators (and the students they are advising) to become affiliated with the lab.

XVI. Outreach and Professional Activities:

Dr. Neill gave guest lectures related to his work on this project in the Heinz Ph.D. seminar, Heinz Faculty Research Seminar, and SCS Immigration Course at CMU. He also led a university-wide grant proposal effort (consisting of over 35 CMU faculty members) to create a Center for Development Data Analytics, focused on applying state-of-the-art machine learning methods to various problems in international development. This proposal is currently under review by the U.S. Agency for International Development (USAID), and at least one piece of this work (focusing on Dr. Neill's event detection and prediction work, as applied to detection of patterns of human rights abuses, violence, and conflict) has already received great attention from the MacArthur Foundation, likely resulting in a funded project within the next year.

As noted above, Dr. Neill will be organizing a CMU workshop on "Machine Learning for the Social Sciences" in October; additionally, he served as Scientific Program Chair for the 2011 International Society for Disease Surveillance Annual Conference. As program chair, his responsibilities included choosing and directing the Scientific Program Committee, coordinating the abstract submission and review process, and inviting speakers, with principal responsibility for the scientific content of the conference. The conference was an outstanding success, with 321 registered participants, 96 contributed talks, 65 posters, and 10 invited talks. Finally, Dr. Neill has been named the AI and Health Department Editor of *IEEE Intelligent Systems*, with the goal of presenting novel research and exciting applications at the intersection of artificial intelligence, machine learning, and health.

References

H. S. Burkom, S. P. Murphy, J. Coberly, and K. Hurt-Mullen, K. Public health monitoring tools for multiple data streams. *Morbidity and Mortality Weekly Report* , 54 (Supplement):55–62, 2005.

K. Das, J. Schneider, and D. B. Neill. Anomaly pattern detection in categorical datasets. *Proc. 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 169-176, 2008.

S. Flaxman and D. B. Neill. Detecting spatially localized subsets of leading indicators for event prediction. Submitted for publication, 2012.

W. R. Hogan, G. F. Cooper, G. L. Wallstrom, M. M. Wagner, and J. M. Depinay. The Bayesian aerosol release detector: an algorithm for detecting and characterizing outbreaks caused by an atmospheric release of Bacillus anthracis. *Statistics in Medicine* 26(29): 5225-52, 2007.

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

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

T. Kumar and D. B. Neill. Fast tensor scan for event detection and characterization. Submitted for publication, 2012.

Y. Liu and D.B. Neill. Detecting previously unseen outbreaks with novel symptom patterns. *Emerging Health Threats Journal* 4: 11074, 2011.

E. McFowland III, S. Speakman, and D.B. Neill. Fast generalized subset scan for anomalous pattern detection. Submitted for publication, 2012.

D. B. Neill. Detection of Spatial and Spatio-Temporal Clusters. Ph.D. thesis, Carnegie Mellon University, School of Computer Science, 2006.

D. B. Neill. Fast Bayesian scan statistics for multivariate event detection and visualization. *Statistics in Medicine* 30(5): 455-469, 2011.

D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360, 2012.

D.B. Neill. New directions in artificial intelligence for public health surveillance. *IEEE Intelligent Systems* 27(1): 56-59, 2012.

D. B. Neill and G. F. Cooper. A multivariate Bayesian scan statistic for early event detection and characterization. *Machine Learning* 79: 261-282, 2010.

D. B. Neill, G. F. Cooper, K. Das, X. Jiang, and J. Schneider. Bayesian network scan statistics for multivariate pattern detection. In J. Glaz, V. Pozdnyakov, and S. Wallenstein, eds., *Scan Statistics: Methods and Applications*, 221-250, 2009.

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

D. B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. Submitted for publication, 2012.

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

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

K. Shao, Y. Liu, and D.B. Neill. A generalized fast subset sums framework for Bayesian event detection. *Proceedings of the 11th IEEE International Conference on Data Mining*, 617-625, 2011.

S. Somanchi and D.B. Neill. Fast graph structure learning from unlabeled data for outbreak detection. *Emerging Health Threats Journal* 4: 11017, 2011.

S. Somanchi and D.B. Neill. Fast graph structure learning from unlabeled data for event detection. Submitted for publication, 2012.

S. Speakman and D. B. Neill. Fast graph scan for scalable detection of arbitrary connected clusters. In *Proc. International Society for Disease Surveillance Annual Conference*, 2010.

S. Speakman, E. McFowland III, and D.B. Neill. Scalable detection of anomalous patterns with connectivity constraints. Submitted for publication, 2012.

S. Speakman, S. Somanchi, E. McFowland III, and D.B. Neill. Scalable detection of anomalous subgraphs. Submitted book chapter, *Encyclopedia of Social Network Analysis and Mining*.

S. Speakman, E. McFowland III, S. Somanchi, and D.B. Neill. Scalable detection of irregular disease clusters using soft compactness constraints. *Emerging Health Threats Journal* 4: 11121, 2011.

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