Results and Findings

I. Theory of linear-time subset scanning

In Neill (2012), we present formal proofs of the linear-time subset scanning property for a large class of score functions, using the "proof by inclusion" and "proof by substitution" methods described in the "Research Activities" section. First, we show (via proof by inclusion), that if F(S) can be written as a function of two additive sufficient statistics $X(S) = \sum_{i\in S} x_i$ and $Y(S) = \sum_{i\in S} y_i$, and if F(X,Y) is monotonically increasing with X, quasi-convex, and restricted to positive Y, then F(S) satisfies the LTSS property. As a corollary, this set of sufficient conditions allows us to prove that Kulldorff's (1997) spatial scan statistic satisfies LTSS, as do the more recently proposed expectation-based Poisson (Neill et al., 2005) and Gaussian (Neill, 2006) statistics. We also present a counter-example which shows that these functions do not satisfy the strong LTSS property.

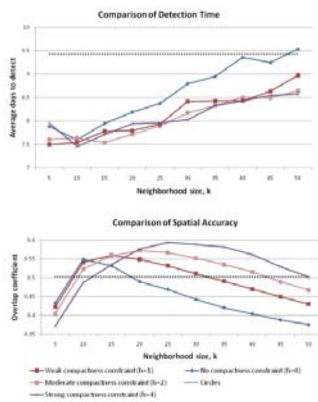
Second, we show (via proof by substitution) that if F(S) can be written as F(X(S), |S|) for a single additive sufficient statistic $X(S) = \sum_{i\in S} x_i$, and if F(X) is monotonically increasing with X, then F(S) satisfies the (strong) LTSS property. As a corollary, this set of sufficient conditions for LTSS allows us to prove that the nonparametric scan statistics (Neill and Lingwall, 2007) satisfy LTSS, as does a simplified version of the Bayesian scan statistic proposed by Neill et al. (2006).

Finally, we show that LTSS holds for a large class of expectation-based scan statistics, likelihood ratio statistics for which the alternative hypothesis $H_1(S)$ assumes a constant multiplicative increase in the expected values in region S. In particular, we show that LTSS holds for any distribution in the separable exponential family, including the expectation-based Poisson, Gaussian, and exponential scan statistics, as well as a novel scan statistic used to detect changes in the variance of a Gaussian distribution with unknown mean.

II. Univariate linear-time subset scanning for spatial datasets

In Neill (2012) we evaluate the run time, detection power, and spatial accuracy of our fast proximity-constrained subset scan approaches, using simulated disease outbreaks injected into real-world hospital Emergency Department data from 97 Allegheny County zip codes. With LTSS, run time increases linearly with neighborhood size, up to a maximum of 0.05 seconds per day of data for the parametric scan methods, and 0.07 seconds for the nonparametric scan. Without LTSS, run time increases exponentially with neighborhood size (k), requiring approximately 30 minutes per day of data at k = 25 and nearly two years for a single day of data at k = 40. Fast multi-scan required an average run time of 0.77 seconds per day of data for the expectation-based Poisson statistic, 0.89 seconds for Kulldorff's spatial scan statistic, 0.88 seconds for the expectation-based Gaussian statistic, and 2.34 seconds for the nonparametric scan statistic. Thus both the fast localized scan and fast multi-scan enable efficient maximization of the score function (detecting the most interesting spatial pattern) for problems that would otherwise be computationally infeasible. Finally, we compared the methods to Kulldorff's original spatial scan method (searching over circular regions). Our results demonstrate that proximity-constrained subset scans substantially improve the timeliness and accuracy of detection, detecting two days faster with fewer than half as many missed outbreaks. More detailed results are provided in Neill (2012).

In Speakman, McFowland, Somanchi, and Neill (2012), the penalized subset scan (assuming soft constraints on spatial compactness) was evaluated using Emergency Department (ED) data from 97 Allegheny County zip codes. We compared detection power and spatial accuracy, with and without the soft constraints, on synthetic, spatially localized outbreaks injected into the ED data. Our results show that including soft compactness constraints allows the penalized subset scan methods to detect outbreaks earlier and improve spatial accuracy, as compared to the unpenalized fast localized scan and circular scan, over a wide range of neighborhood sizes (see figure below).

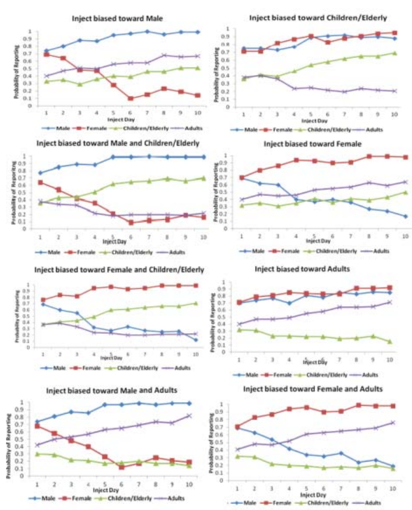


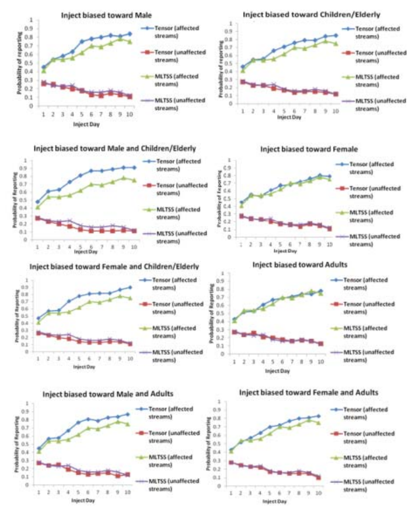
III. Linear-time subset scanning for multivariate and tensor datasets

In Neill, McFowland, and Zheng (2012) we evaluated the speed of the LTSS-enabled multivariate scan statistic methods, demonstrating huge improvements over the naïve approach. For the Subset Aggregation method, our FN method scales linearly rather than exponentially in the number of spatial locations, thus efficiently optimizing the score function when the number of streams is small. Similarly, our NF method scales linearly rather than exponentially in the number of streams, thus efficiently optimizing the score function when the number of search regions (subsets of locations) considered is small. When both the number of streams and number of locations are large, our FF method must be used. Though FF converges to a local maximum and is not guaranteed to obtain the global maximum, we demonstrate that it is able to find regions scoring very close to the global maximum in practice. When monitoring 16 public health data streams over 97 Allegheny County zip codes, the FF method achieves a 200x speedup as compared to FN, while achieving a score within 1% of the global maximum approximately 99% of the time. For the Kulldorff method, our FK method scales linearly rather than exponentially in the number of locations, making this pattern detection method computationally feasible when the number of locations is large. When monitoring 16 public health data streams over 97 Allegheny County zip codes, the FK method achieves a score within 5% of the global maximum approximately 95% of the time.

We then compared the detection performance of the Subset Aggregation and Kulldorff variants of the multivariate spatial scan for synthetic and real-world disease surveillance datasets. For both methods, we compared our "fast localized scan" approach (searching over all subsets of locations constrained by spatial proximity) to the traditional spatial scan approach (searching over circular regions). We demonstrated that the fast localized scan significantly improved detection power and spatial accuracy for both Subset Aggregation and Kulldorff methods, while maintaining efficient and scalable computation. Comparing the Subset Aggregation and Kulldorff methods, we demonstrated that Kulldorff's method tends to achieve somewhat higher detection power and spatial accuracy when data streams are affected to differing extents. However, when only a subset of streams is affected, Subset Aggregation more accurately characterizes events by identifying the affected streams. Thus Kulldorff's method may be preferable when event detection is the primary goal, while Subset Aggregation may be preferable when event characterization is paramount. Our fast, scalable algorithms enable either method to be effectively and efficiently applied to massive, high-dimensional datasets.

We evaluated the Fast Tensor Scan using simulated disease outbreaks injected into real-world Emergency Department data form Allegheny County, PA. In addition to demonstrating substantial improvements in detection time and spatial accuracy for outbreaks which have differing impacts on different subpopulations, FTS was able to accurately identify the characteristics (including, but not limited to, demographics such as age, gender, race/ethnicity, and socioeconomic status) of the affected subpopulation. Additionally, detection performance was further enhanced by incorporating additional constraints such as spatial proximity and graph connectivity into the iterative detection procedure.

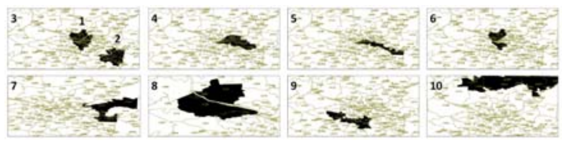




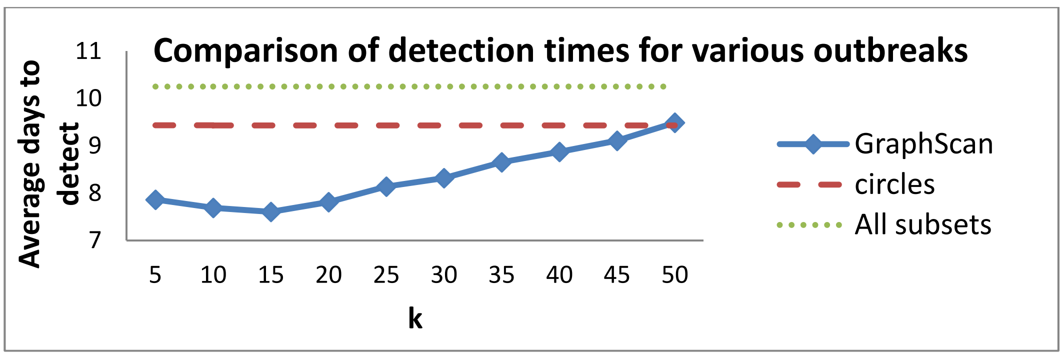
For example, the set of graphs on the left (Figure 3 from Kumar and Neill, 2012) demonstrates that FTS is able to correctly report the affected demographic groups, while the set of graphs on the right (Figure 2 from Kumar and Neill, 2012) demonstrates that FTS outperforms our previous multivariate LTSS approach in terms of correctly reporting the three affected data streams (diarrheal, respiratory, and fever Emergency Department visits) without reporting the unaffected streams.

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

We have applied GraphScan to the spatial event detection domain of biosurveillance. We performed empirical tests of GraphScan's detection time and spatial accuracy using a set of simulated respiratory disease outbreaks injected into real-world Emergency Department data from Allegheny County, Pennsylvania. We compare results for three different methods; ''circles'' (traditional approach introduced by Kulldorff, returns the highest scoring circular cluster of locations), “all subsets” (LTSS implemented without proximity or connectivity constraints, returns the highest scoring unconstrained subset of locations), and “GraphScan” (returns the highest scoring connected subset within a neighborhood size of *k*). The simulated injects are increasing in severity over the course of 14 days, and we compare the methods’ detection times at a fixed false positive rate of 1 per month.



The experiments were performed on 10 different types of spatial injects ranging from compact clusters, irregularly shaped elongated clusters, and even disjoint ones, as shown above. Two hundred outbreaks of each type were evaluated, and the aggregate results for all 2,000 outbreaks are shown below.



GraphScan was able to detect the outbreaks 1.8 days earlier on average compared to the traditional method of scanning over circular regions. The results from the unconstrained “all subsets” method are poor, demonstrating the importance of using proximity and connectivity to constrain the search. GraphScan also outperformed the “circles” and “all subsets” approaches in terms of spatial accuracy, evaluated by measuring the average overlap coefficient between the true and detected regions.

The figure below (Figure 7 from Speakman, McFowland, and Neill, 2012) shows GraphScan’s huge increase in speed over previous methods that searched for the highest scoring connected subset. The previously proposed FlexScan method (Tango and Takahashi, 2005) required a week of computation time to search over neighborhoods of size 30 for a single day’s worth of data. Through the speed improvements of the LTSS property, GraphScan is able to perform this search in less than one second. Searching over all connected subgraphs (without neighborhood constraints) for the 97 Allegheny County zip codes requires about 2.5 seconds for the latest version of GraphScan (as compared to about 2 minutes for the previous version) and would require hundreds of millions of years for FlexScan.

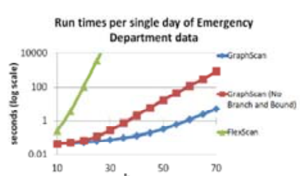


Figure 7: Run time analysis for FlexScan and GraphScan with and without Branch and Bounding. The x-axis denotes the “neighborhood size” as various values of k.

V. Using GraphScan to learn the underlying graph structure

We demonstrated both theoretically and empirically that our graph learning framework requires fewer calls to BestSubgraph than a naive greedy approach, $O(N^3)$ as compared to $O(N^4)$ for exact greedy search, and $O(N \log N)$ as compared to $O(N^2)$ for approximate greedy search, resulting in 2 to 3 orders of magnitude speedup in practice. All five methods were able to learn the structure of graphs with 100 nodes, typically requiring less than an hour of run time, and the faster methods (using the ULS algorithm for event detection) were able to scale to graphs with 500-1000 nodes. It is still an open question whether these methods can be extended to scale up to massive graphs with millions of nodes, as necessary for evaluation of online social network data, and we hope to address this in future work.

We tested the detection power and spatial accuracy of these approaches for detecting 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), and observed tradeoffs between scalability and detection performance.

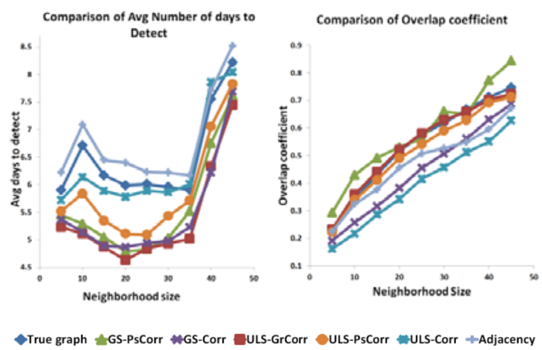


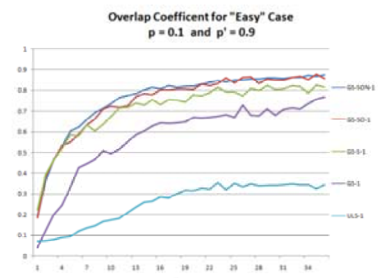
Figure 3 from Somanchi and Neill (2012): Comparison of detection performance of the true, learned, and adjacency graphs for injects based on adjacency with simulated travel patterns.

We demonstrated that the graph structures learned by our framework can then be used by graph-based event detection methods such as GraphScan, enabling more timely and more accurate detection of events (such as disease outbreaks) which spread based on that latent structure. For example, the above figure compares performance of the five methods on detecting outbreaks which spread via spatial adjacency plus (simulated) travel patterns, demonstrating improved detection power and spatial accuracy, as compared to the true graph (spatial adjacency + travel) and the spatial adjacency graph without travel patterns.

In general, the learned graph structures tend to be similar to the true underlying graph structure, capturing nearly all of the true edges but also adding some additional edges. The resulting graph achieves similar spatial accuracy to the true graph, as measured by the overlap coefficient between true and detected clusters. Interestingly, the learned graph often has <u>better</u> detection power than the true underlying graph, enabling more timely detection of outbreaks or other emerging events. This result can be better understood when we realize that 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 timely and accurate event detection.

VI. Detecting dynamic events with temporal consistency constraints

In Speakman and Neill (in preparation), we empirically show that the incorporation of temporal consistency constraints into our GraphScan algorithm provides substantial improvements in detecting and tracing contaminants spreading through a water distribution system. The data comes from the Battle of the Water Sensor Networks project and has typically been used to study the optimal placement of perfect sensors. We do not address the sensor placement problem in this work, but instead explore the time to detection and spatial accuracy of various graph-based detection methods when imperfect sensors are used in the system. Our results show that incorporating simple size and temporal consistency constraints allows GraphScan to detect the contaminants earlier and more accurately identify which nodes are affected as the contamination spreads through the network.



For example, the above figure compares the spatial accuracy (overlap coefficient) of five graph-based detection methods for the "easy" case where the sensors' false positive and false negative rates are set to 10% (the full paper also includes similar results for the "medium" and "hard" cases of 20% and 40% respectively). All three variants of penalized GraphScan outperform the original version of GraphScan (purple), which required an average of 9.65 hours to detect the contamination, and Patil and Taillie's Upper Level Set Scan Statistic (turquoise), which performed extremely poorly (15.4 hours to detect). Using a penalty on the size of the detected subset (green) improves spatial accuracy as compared to the unpenalized GraphScan, but does not detect any faster (10.27 hours to detect). However, including temporal consistency constraints decreases the detection time to 7.82 and 7.66 hours for the GS-SO-1 method (rewarding nodes which were part of the best subset on the previous two time steps) and the GS-SON-1 method (rewarding nodes which were part of the best subset on the previous two time steps, and the neighbors of these nodes) respectively. Overall, enforcing temporal consistency constraints improved detection time by two hours when compared to the unpenalized version of GraphScan.

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

In McFowland, Speakman, and Neill (2012, submitted for publication), we evaluated the ability of our Fast Generalized Subset Scan (FGSS) method to detect anomalies in general datasets from three real-world domains. First, we considered the 1999 KDDCup network intrusion dataset, which contains a variety of simulated network intrusions on a military network environment. Second, we considered synthetic anomalies injected into real container shipment data, simulating the anomalous patterns which may result from illicit activities such as smuggling or terrorism. Finally, using a state-of-the-art simulator (Hogan et al., 2007), we simulated the effects of an airborne anthrax release on the number and spatial distribution of respiratory ED cases. We compared FGSS to several previously proposed methods: Anomaly Pattern Detection (Das et al., 2008), Anomalous Group Detection (Neill et al., 2009), and the Bayesian Network Anomaly detector (which identifies individually anomalous records). We evaluated each method's ability to a) distinguish between anomalous and normal records in a dataset which contains anomalies (as measured by the precision vs. recall curve), b) distinguish between datasets containing anomalies and those without anomalies (as measured by the ROC curve), c) identify which attributes are anomalous for the affected group of records (as measured by the overlap coefficient between true and reported sets of attributes), and d) scale to large and high-dimensional data (as measured by runtime as a function of the number of records N and number of attributes M). Our results demonstrate that FGSS can successfully detect and characterize useful anomalous patterns in all three of these application domains, and can scale up to huge and high-dimensional datasets. For example, the following table (Table 1 from McFowland, Speakman, and Neill, 2012) demonstrates substantial improvements in performance for the container shipment dataset as measured by area under the precision vs. recall curve.

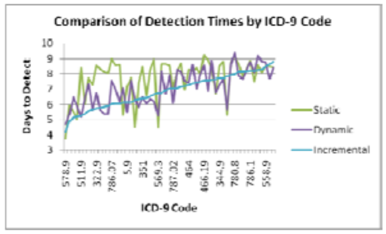
Table 1: PIERS Container Shipment Data: Average area (in percent) under the PR curve, with standard errors. For each row, the method which demonstrates the best performance, and those methods with performance not significantly different at significance level $\alpha = 0.05$, are bolded.

N	k_{inj}	s_{inj}	m_{inj}	$FGSS-BJ$	$FGSS-HC$	AGD	APD	BN
1000	1	10	1	**76.9±3.9**	52.3±4.7	62.2±4.2	47.7±4.3	18.8±2.7
1000	1	10	2	**80.9±3.2**	67.6±4.0	64.9±4.1	65.5±4.1	38.9±3.7
1000	4	25	1	**94.2±1.0**	61.7±2.9	**93.0±1.2**	52.9±2.0	43.5±2.2
1000	4	25	2	**97.3±1.0**	87.3±1.6	94.3±0.7	77.3±1.5	73.1±1.8
1000	10	10	1	**90.8±1.2**	62.0±1.9	80.4±1.5	52.9±1.6	39.6±1.5
1000	10	10	2	**91.5±0.8**	85.4±1.0	83.5±1.0	75.7±1.3	71.4±1.2
10,000	4	25	1	**79.7±2.5**	40.2±3.5	These	41.8±2.9	8.0±1.0
10,000	4	25	2	**71.2±1.8**	65.2±2.5	runs	64.2±2.6	26.4±1.9
10,000	10	10	1	**54.3±2.4**	41.5±2.7	did not	16.2±1.4	6.8±0.8
10,000	10	10	2	51.6±2.0	**65.4±1.7**	complete	40.2±2.3	26.7±1.4

In general, FGSS consistently outperformed APD and BN with respect to detection power, with comparable or better performance as compared to AGD. Additionally, FGSS was able to scale to datasets consisting of 100,000 data records, while AGD fails to scale even to 10,000 records, and FGSS was also able to much more accurately characterize the affected subset of attributes as compared to the competing methods.

VIII. Semantic scan statistic

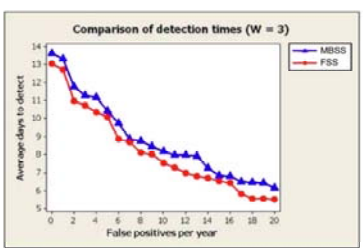
In Liu and Neill (2011), we compared the performance of the semantic scan statistic to the traditional spatial scan approach on the outbreak detection task, using Emergency Department data from Allegheny County, PA. Our data consisted of 612,713 cases from 1/1/2004-12/31/2005, with a free-text chief complaint, International Classification of Diseases (ICD-9) code, and general syndrome classification (prodrome) for each case. We generated synthetic outbreaks consisting of cases drawn at random using a given ICD-9 code, and compared the results of the standard spatial scan (maximizing over the ten prodrome categories: respiratory, gastrointestinal, etc.) and the semantic scan (generating topics from the free text and maximizing over the topics). We used average detection time (in days) as well as proportion of outbreaks detected, at a fixed false positive rate of 1 per month, as our evaluation metrics. The figure below compares the detection performance of the static, dynamic, and incremental subset scan methods, demonstrating that the dynamic and incremental methods tend to achieve more timely detection than the static method across a wide range of disease syndromes (as grouped by International Classification of Diseases (ICD)-9 codes).

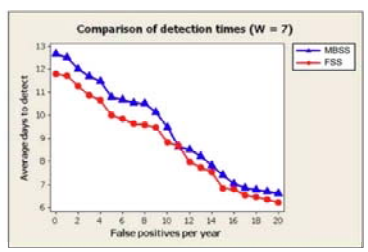


Additionally, all three semantic scan methods were able to accurately identify synthetically-generated unprecedented outbreaks (e.g. a disease which makes the patient's nose turn green), detecting in an average of 5.3 to 6.4 days, while the typical prodrome-based method had low power to detect such novel outbreaks, requiring an average of 10.9 days to detect.

IX. Bayesian fast subset sums

In Neill (2011), we compare the run time, detection power, and spatial accuracy of our Bayesian fast subset sums (FSS) method to our original multivariate Bayesian scan statistic (MBSS) approach. We used simulated irregularly-shaped disease outbreaks injected into real-world Emergency Department data from Allegheny County, PA. While the run time of FSS was approximately 7.5× slower than the original MBSS method, it is still extremely fast, computing the posterior probability map for each day of data in less than nine seconds.

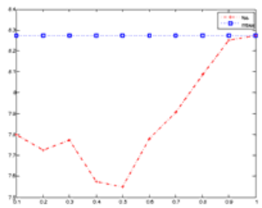


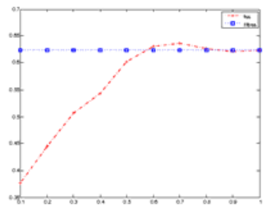


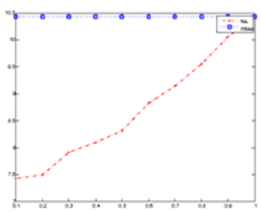
As shown above, FSS was able to detect outbreaks an average of one day earlier (for a fixed false positive rate of 1/month) as compared to the original MBSS approach for a maximum temporal window size of W = 3, and detected an average of 0.54 days earlier for a maximum temporal window size of W = 7. As shown in the example below, FSS was also better able to determine the affected spatial region for irregularly-shaped clusters, showing significantly improved spatial accuracy (overlap coefficient) as compared to MBSS. Complete results are presented in Neill (2011).

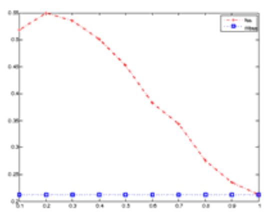
Highly elongated cluster, downtown Pittsburgh. Left: true outbreak region.
Center: average posterior probability map for MBSS. Right: average posterior probability map for FSS.

Finally, we have developed a generalization of FSS which also allows another parameter (sparsity of the detected region) to be controlled. For a given neighborhood (consisting of a center location and its k-nearest neighbors), whether each location is affected is assumed to be drawn independently with probability p. The original FSS approach (which assumes that all subsets are equally likely given the neighborhood) corresponds to p = 0.5, and the MBSS approach (which searches over circular-shaped regions only) corresponds to p = 1. However, our recent results demonstrate that better choices of the p parameter lead to improved detection time and spatial accuracy.

a) 

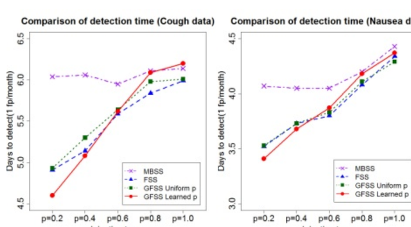
b) 

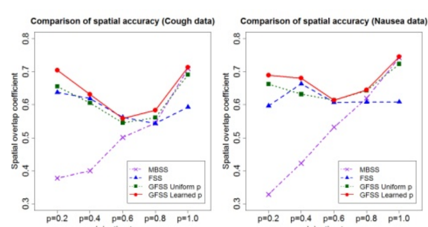
c) 

d) 

In this figure, a) and b) show the detection time and spatial accuracy for a compact outbreak respectively: detection time is minimized at p = 0.5, while spatial accuracy is maximized at p = 0.7. Similarly, c) and d) show the detection time and spatial accuracy for a highly elongated outbreak respectively: detection time is minimized at p = 0.1, while spatial accuracy is maximized at p = 0.2. These results demonstrate that optimization of the sparsity parameter can substantially improve the detection performance of the generalized FSS approach.

In Shao, Liu, and Neill (2011), we demonstrate that the Generalized Fast Subset Sums (GFSS) method with learned distribution of the sparsity parameter p outperforms our previously proposed Multivariate Bayesian Scan Statistic (MBSS) and Fast Subset Sums (FSS) methods, as well as the GFSS method assuming a uniform distribution of p. The figures below (Figures 7 and 8 of Shao, Liu, and Neill, 2011) compare these methods with respect to detection power and spatial accuracy respectively, demonstrating the value of learning the model parameters from data.





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

In Flaxman and Neill (2012, submitted), we demonstrated that our Iterative Average Dot Product (IADP) method can accurately maximize cross-correlation over subsets of the monitored data streams and proximity-constrained subsets of locations. For jointly maximizing cross-correlation over streams and locations, we compared to ground truth for small problem sizes (for which we can compute the true maximum by exhaustive search over all subsets) and performed exploratory analysis for large subsets. For maximizing cross-correlation over streams for a given set of locations, we also compared to previous methods in the literature (LASSO and Google Correlate). Our results demonstrate that IADP has similar run time to Google Correlate and much faster run time than LASSO, while finding subsets with significantly higher correlation than either competing method. For our exploratory analysis, IADP found a spatial area consisting of 10 census tracts in the West Englewood neighborhood of Chicago, and a subset of 12 of the 28 monitored data streams, for which the correlation between this week's leading indicator count and next week's violent crime count was 0.786 (time series shown below); the average correlation of the regions it found, averaged over all 77 community areas, was 0.546.

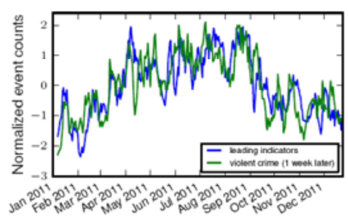


Figure 3: Best correlation discovered in table 6 between LIs and VC is $r = 0.786$, occurring in community area 67 (West Englewood).

In comparison, the LASSO and Google Correlate methods (which could not search over subsets of census tracts, and thus simply aggregated counts for each community area), found much lower maximum correlations (0.660 and 0.652 for West Englewood respectively) and much lower average correlations (0.325 and 0.404 respectively).

XIV-XV. Results of Educational Initiatives

As described in detail in "Research and Educational Activities", this research project has directly resulted in the mentoring of twelve graduate students (seven current, five graduated), including the completion of four First and Second Heinz Research Papers (graduation requirements for the Ph.D. in Heinz College at CMU), one master's thesis, and two capstone projects. Additionally, Dr. Neill has developed a new curriculum in Machine Learning and Policy at CMU, including three new courses (described in "Research and Educational Activities") and a new joint Ph.D. program in Machine Learning and Public Policy, with the first student joining the program last fall. The special topics course, "Special Topics in Machine Learning and Policy: Harnessing the Wisdom of Crowds", was taught for the first time in Spring 2012, with high relevance to future work incorporating crowdsourced data on this project. Last year's special topics course, "Machine Learning for the Developing World", helped to connect the PI and project team with other international development researchers on campus, leading to a university-wide effort led by the PI to establish a Center for Development Data Analytics.

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.