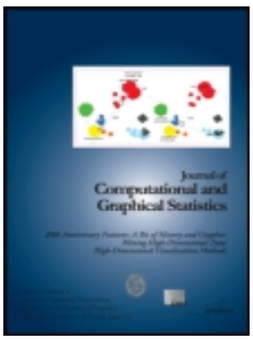

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: http://www.tandfonline.com/loi/ucgs20

Penalized Fast Subset Scanning

Skyler Speakman, Sriram Somanchi, Edward McFowland III & Daniel B. Neill

To cite this article: Skyler Speakman, Sriram Somanchi, Edward McFowland III & Daniel B. Neill (2016) Penalized Fast Subset Scanning, Journal of Computational and Graphical Statistics, 25:2, 382-404, DOI: 10.1080/10618600.2015.1029578

To link to this article: http://dx.doi.org/10.1080/10618600.2015.1029578

Accepted author version posted online: 18 Apr 2015.
Published online: 10 May 2016.

Penalized Fast Subset Scanning

Skyler SPEAKMAN, Sriram SOMANCHI, Edward MCFOWLAND III, and Daniel B. NEILL

We present the penalized fast subset scan (PFSS), a new and general framework for scalable and accurate pattern detection. PFSS enables exact and efficient identification of the most anomalous subsets of the data, as measured by a likelihood ratio scan statistic. However, PFSS also allows incorporation of prior information about each data element's probability of inclusion, which was not previously possible within the subset scan framework. PFSS builds on two main results: first, we prove that a large class of likelihood ratio statistics satisfy a property that allows additional, element-specific penalty terms to be included while maintaining efficient computation. Second, we prove that the penalized statistic can be maximized exactly by evaluating only O(N) subsets. As a concrete example of the PFSS framework, we incorporate "soft" constraints on spatial proximity into the spatial event detection task, enabling more accurate detection of irregularly shaped spatial clusters of varying sparsity. To do so, we develop a distance-based penalty function that rewards spatial compactness and penalizes spatially dispersed clusters. This approach was evaluated on the task of detecting simulated anthrax bio-attacks, using real-world Emergency Department data from a major U.S. city. PFSS demonstrated increased detection power and spatial accuracy as compared to competing methods while maintaining efficient computation.

Key Words: Disease surveillance; Likelihood ratio statistic; Pattern detection; Scan statistic.

1. INTRODUCTION

Detecting patterns in massive datasets has multiple real-world applications in fields such as public health, law enforcement, and security. For example, spatial scan statistics are commonly used to alert public health officials to an unexpected increase in the number of Emergency Department complaints from patients in some spatial region (i.e., set of nearby zip codes), which may indicate the early stages of an emerging disease outbreak or attack.

In this work, we consider the "subset scan" approach to pattern detection, which treats the problem as a constrained search over subsets of data elements, with the goal of finding the most anomalous subsets. Unlike "bottom-up" approaches that find and aggregate individual

Skyler Speakman (E-mail: *skylerspeakman@gmail.com*), Sriram Somanchi (E-mail: *ssv.sriram@gmail.com*), Edward McFowland III (E-mail: *emcfowla@umn.edu*), and Daniel B. Neill (E-mail: *neill@cs.cmu.edu*), Associate Professor and Director, Event and Pattern Detection Laboratory, Carnegie Mellon University, Pittsburgh, PA 15213.

Journal of Computational and Graphical Statistics, Volume 25, Number 2, Pages 382–404
DOI: 10.1080/10618600.2015.1029578
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jcgs*.

anomalies, and "top-down" approaches that detect globally anomalous trends and then localize them, subset scanning approaches maintain high detection power for both highly localized and global trends (Neill 2009b, 2012). However, subset scanning approaches pose two main challenges. First is appropriately evaluating the "anomalousness" of a given subset, and second is the computational issue of searching through the 2^N possible subsets of a dataset containing N elements. Previous approaches (Kulldorff 1997; Neill et al. 2005; Neill 2009b, 2012) have addressed the first concern by "scoring" each subset using a log-likelihood ratio statistic, such as the expectation-based scan statistics (Neill et al. 2005; Neill 2009b) considered here.

For spatial data, the computational challenge of subset scanning has been addressed in several ways: limiting the search space to only consider regions of a given shape, such as circles (Kulldorff 1997) or rectangles (Neill and Moore 2004; Wu et al. 2009), or performing a heuristic search over subsets, which is not guaranteed to find the most anomalous subsets (Duczmal and Assuncao 2004; Agarwal et al. 2006). Such approaches enable efficient computation at the expense of reduced detection power and spatial accuracy (Neill 2012). The fast subset scan (FSS) approach (Neill 2012) resolves these computational challenges through exact and efficient identification of the highest-scoring subset, for any score function satisfying the linear-time subset scanning (LTSS) property. However, rather than performing an unconstrained search over all possible subsets of the data, we often wish to incorporate either "hard constraints" (ruling out some subsets completely) or "soft constraints" (penalizing less likely subsets) into the search procedure.

Certain types of hard constraints are possible to incorporate within the FSS framework: for example, the "fast localized scan" (Neill 2012) enforces a hard constraint on spatial proximity by performing a search over the "local neighborhood" consisting of each spatial location and its $k - 1$ nearest neighbors. Similarly, GraphScan (Speakman, McFowland, and Neill 2015) incorporates hard connectivity constraints by ruling out subsets that are disconnected in an assumed underlying graph structure. However, soft constraints (e.g., a prior belief that some locations are more likely to be affected) cannot be easily incorporated. Given a score function satisfying the LTSS property, a penalized version of that score function is not guaranteed to satisfy LTSS, and thus FSS cannot efficiently identify the highest-scoring *penalized* subset. An example is provided in Section 4.

In this work, we introduce and formalize a new property of scoring functions, Additive linear-time subset scanning (ALTSS), which allows incorporation of prior information about each data element's probability of inclusion. We demonstrate that many commonly used log-likelihood ratio scan statistics satisfy the ALTSS property. Thus, we show that the penalized version of these statistics (where we have included each element's prior log-odds of being part of an anomalous subset as a bonus or penalty for including that element) can be exactly and efficiently optimized over all subsets of the data, without requiring an exhaustive search over all subsets.

We highlight three contributions in this work that follow from the ALTSS property. The first is the penalized fast subset scan (PFSS) framework laid out in Section 3. PFSS is very general, enabling any element-specific priors to be incorporated into the search over subsets while maintaining computational efficiency and exactness. Our second contribution is an investigation of the connections between ALTSS and the LTSS property (Neill 2012).

More specifically, we show that scoring functions in the form of expectation-based scan statistics from the exponential family satisfy LTSS. This contribution extends LTSS, which was previously limited to the "separable" subfamily of the exponential family. Expectation-based scan statistics using the binomial and negative binomial distributions (which are not part of the separable subfamily) may now be efficiently optimized in their penalized and unpenalized forms.

Our final contribution is a specific application of PFSS based on motivating examples from the fields of bio-terrorism and disease surveillance. While the "fast localized scan" (subset scan with hard constraints on spatial proximity) has been shown to achieve high detection power and spatial accuracy in this setting (Neill 2012), it does not take into account the spatial attributes of the locations beyond the "hard" proximity constraint of being one of the $k-1$ nearest neighbors of a center location, and considers each of the 2^k subsets of the neighborhood equally likely.

Soft proximity constraints incorporate the prior expectation that locations closer to the center of an outbreak are more likely to be affected, thus rewarding spatial compactness and penalizing spatially dispersed clusters. We demonstrate that this approach increases both detection power and spatial accuracy as compared to the fast localized scan. Additionally, while fast localized scan achieves high performance for well-chosen values of the neighborhood size k, it performs worse than the standard, circular spatial scan (Kulldorff 1997) for badly chosen k. We demonstrate in Section 7 that incorporation of soft constraints enables our penalized version of the fast localized scan to be much more robust to the choice of k, while still guaranteeing that the most anomalous *penalized* subset of locations will be exactly and efficiently identified. This robustness to parameter selection is critical when a limited number of labeled training examples exist or when a public health surveillance system must be able to detect a wide range of possible outbreak types.

1.1 EXPECTATION-BASED SCAN STATISTICS

We now review the use of expectation-based scan statistics (Neill et al. 2005) for spatial event detection. In the subset scanning framework, our goal is to identify a subset of the data $S \subseteq D$ that maximizes a score function $F(S)$. In the spatial event detection setting considered here, the dataset D consists of spatial time series data: observed counts x_i and expected counts μ_i at a set of spatial locations s_i $(i = 1, \ldots, N)$ and possibly other parameters, such as the standard deviations σ_i. Likelihood ratio statistics have been commonly used as score functions (Kulldorff 1997; Neill et al. 2005). The log-likelihood ratio statistic is defined as $F(S) = \log(\Pr(D \mid H_1(S))/\Pr(D \mid H_0))$, where the alternative hypothesis $H_1(S)$ assumes an event occurring in region $S \subseteq \{s_1, s_2, \ldots, s_N\}$ and the null hypothesis H_0 assumes that no events are occurring. For the expectation-based scan statistics, the alternative hypothesis $H_1(S)$ assumes that counts x_i are drawn with mean $q\mu_i$ inside region S and mean μ_i outside region S, for some constant multiplicative factor $q > 1$ known as the *relative risk* or severity. We can then write the expectation-based scan statistic as

$$F(S) = \max_{q>1} \sum_{s_i \in S} (\log \Pr(x_i \mid q\mu_i) - \log \Pr(x_i \mid \mu_i)). \quad (1)$$

A pivotal insight of our work is that for a *fixed* value of the relative risk q, the expectation-based scan statistics from the exponential family can be written as an *additive* set function over the data elements s_i contained in S. This insight leads to three useful consequences. First, additional penalty terms may be added *at the element level* (i.e., a bonus or penalty Δ_i for each element s_i) and the resulting penalized function will still be additive. Second, the highest scoring penalized subset can be efficiently identified by selecting only those data elements s_i making a positive contribution to the penalized scoring function. Finally, we show in Section 3.2 that only a small number of values of q must be considered, thus leading to efficient optimization of (penalized or unpenalized) score functions $F(S)$ over all $q > 1$.

Kulldorff's spatial scan statistic (Kulldorff 1997) is also a scoring function based on likelihood ratio statistics, but it is not an *expectation-based* scan statistic. Kulldorff's statistic requires two parameters under the alternative hypothesis, q_{in} and q_{out}, which represent the multiplicative increase in counts for locations inside and outside of S, respectively; in this case, $F(S \mid q_{\text{in}}, q_{\text{out}})$ can be written as an additive function. Neill (2009a) demonstrated that Kulldorff's statistic has low detection power for large outbreaks that cover much of the search region, since $q_{\text{in}} \approx q_{\text{out}}$ for all S and no subset appears particularly anomalous. Expectation-based scan statistics use only data from within S and therefore represent a more natural model when identifying locations with higher activity than expected. The expected counts μ_i for expectation-based scan statistics can be derived from a variety of time series forecasting methods, including simple moving averages or more complex functions that adjust for seasonal and day-of-week trends (Burkom, Murphy, and Shmueli 2007).

We note that the assumptions of conditionally independent counts and a constant, multiplicative risk q are standard in the spatial scan literature (Kulldorff 1997). The assumption of constant risk q, estimated by maximum likelihood, is preferable to the alternative of estimating risks q_i independently for each location. The latter approach tends to overfit the noise in the data, since any locations with x_i even slightly larger than μ_i would make a positive contribution to the score function and would be included in the highest-scoring subset.

2. THE ADDITIVE LINEAR TIME SUBSET SCANNING PROPERTY

We now define the ALTSS property. Informally, a score function $F(S)$ satisfies ALTSS if conditioning on the relative risk q allows the function to be written as an additive set function over the data elements s_i contained in S.

Definition 1. For a given dataset D, the score function $F(S)$ satisfies the ALTSS property if for all subsets $S \subseteq D$, we have $F(S) = \max_{q>1} F(S \mid q)$, where $F(S \mid q) = \sum_{s_i \in S} \lambda_i(q)$, and $\lambda_i(q)$ depends only on the given value of q, the observed count x_i, and expected count μ_i (and in some cases standard deviation σ_i) for element s_i.

Theorem 1. Expectation-based scan statistics from the (single parameter) exponential family satisfy the ALTSS property.

Table 1. Derivation of $\lambda_i(q)$ for expectation-based scan statistics in the exponential family

Distribution	$\theta(q\mu_i)$	$\phi(q\mu_i)$	$\lambda_i(q)$
Poisson	$\log(q\mu_i)$	$q\mu_i \log(q\mu_i) - q\mu_i$	$x_i \log q + \mu_i(1-q)$
Gaussian	$\frac{q\mu_i}{\sigma_i^2}$	$\frac{(q\mu_i)^2}{2\sigma_i^2}$	$x_i\mu_i\frac{(q-1)}{\sigma_i^2} + \mu_i^2\left(\frac{1-q^2}{2\sigma_i^2}\right)$
Exponential	$-\frac{1}{q\mu_i}$	$-\log(q\mu_i)$	$\frac{x_i}{\mu_i}(1-\frac{1}{q}) - \log q$
Binomial	$\log\left(\frac{q\mu_i}{n_i - q\mu_i}\right)$	$q\mu_i \log\left(\frac{q\mu_i}{n_i - q\mu_i}\right) + n_i \log(n_i - q\mu_i)$	$x_i \log(q) + (n_i - x_i)\log\left(\frac{n_i - q\mu_i}{n_i - \mu_i}\right)$
Negative binomial	$\log\left(\frac{q\mu_i}{r_i + q\mu_i}\right)$	$q\mu_i \log\left(\frac{q\mu_i}{r_i + q\mu_i}\right) - r_i \log(r_i + q\mu_i)$	$x_i \log(q) + (r_i + x_i)\log\left(\frac{r_i + \mu_i}{r_i + q\mu_i}\right)$

Proof of Theorem 1. Following the notation in Neill (2012), we write the distributions from the exponential family as $\log \Pr(x \mid \mu) = T(x)\theta(\mu) - \psi(\theta(\mu)) = T(x)\theta(\mu) - \mu\theta(\mu) + \phi(\mu)$, where $T(x)$ is the sufficient statistic, $\theta(\mu)$ is a function mapping the mean μ to the natural parameter θ, ψ is the log-partition function, and ϕ is the convex conjugate of ψ. Plugging this form of the exponential family into (1) gives

$$F(S) = \max_{q>1} \sum_{s_i \in S} (T(x_i)(\theta(q\mu_i) - \theta(\mu_i)) + \mu_i\theta(\mu_i) - q\mu_i\theta(q\mu_i) + \phi(q\mu_i) - \phi(\mu_i)). \quad (2)$$

Let $\lambda_i(q) = T(x_i)(\theta(q\mu_i) - \theta(\mu_i)) + \mu_i\theta(\mu_i) - q\mu_i\theta(q\mu_i) + \phi(q\mu_i) - \phi(\mu_i)$ and then $F(S)$ satisfies the ALTSS property.

Table 1 summarizes the derivation of $\lambda_i(q)$ for the expectation-based scan statistics from several distributions in the exponential family.

An important consequence of scoring functions being written as additive functions over the data elements contained in the subset is that additional bonus or penalty terms Δ_i may be included for each data element s_i while maintaining the additive property.

Corollary 1. Given a scoring function $F(S)$ that satisfies the ALTSS property, assume an additive bonus or penalty Δ_i for each $s_i \in S$. The resulting penalized score function,
$F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$, also satisfies ALTSS.

Proof of Corollary 1.

$$F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$$

$$= \max_{q>1} F(S \mid q) + \sum_{s_i \in S} \Delta_i$$

$$= \max_{q>1} \sum_{s_i \in S} (\lambda_i(q) + \Delta_i)$$

$$= \max_{q>1} \sum_{s_i \in S} \gamma_i(q),$$

where $\gamma_i(q) = \lambda_i(q) + \Delta_i$ is referred to as the total contribution of data element s_i to the penalized scoring function for a fixed risk q.

The Δ_i terms are a function of the given data element s_i; they cannot depend on the subset S. We plan to investigate more sophisticated penalties in future work.

A second important consequence of scoring functions being written as additive functions is that the highest scoring subset for a fixed risk q can be easily identified.

Corollary 2. For a fixed risk q, functions satisfying ALTSS can be efficiently optimized over all subsets $S \subseteq D$ by including all and only those data elements making a positive contribution to the scoring function, that is, $s_i \in \arg\max_{S \subseteq D} F(S \mid q)$ if and only if $\gamma_i(q) = \lambda_i(q) + \Delta_i > 0$.

The proof of Corollary 2 follows immediately from the fact that $F(S \mid q) = \sum_{s_i \in S} \gamma_i(q)$.

3. PENALIZED FAST SUBSET SCANNING

Penalized fast subset scanning (PFSS) is a novel method for scalable and accurate pattern detection, which uses the ALTSS property of commonly used scoring functions to incorporate prior information for each data element. This is in contrast to the FSS method (Neill 2012), which does not allow for additional terms to influence the subset's score and therefore considers each element equally likely to be included in the highest scoring *un*penalized subset. The first half of this section focuses on how the additional, element-specific terms are interpreted in the PFSS framework and the second half explains how the penalized scoring function may be exactly and efficiently optimized over all possible subsets.

3.1 PRIOR LOG-ODDS INTERPRETATION OF PENALTIES Δ_i

We first show that the penalty terms Δ_i can be usefully interpreted as the prior log-odds that each data record s_i is affected. Let us assume a simple generative model where some subset of records $S_{\text{true}} \subseteq \{s_1, s_2, \ldots, s_N\}$ is affected, and each s_i is independently chosen to be included in S_{true} with prior probability p_i. We now consider the penalized score function $F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$, where the log-likelihood ratio $F(S) = \log(\Pr(D \mid H_1(S))/\Pr(D \mid H_0))$ and $\Delta_i = \log(p_i/(1 - p_i))$. Given the priors p_i, we show that this choice of Δ_i satisfies two useful properties: the highest-scoring penalized subset $S^* = \arg\max_S F_{\text{pen}}(S)$ minimizes the total probability of error, and is also a maximum a posteriori (MAP) estimate of the true affected subset S_{true}.

First, when comparing the detected subset S^* and the true affected subset S_{true}, we wish to minimize both the probability of incorrectly including extra records (Type I error) and

the probability of failing to detect truly affected records (Type II error). We show that the choice of $\Delta_i = \log(p_i/(1-p_i))$ minimizes the sum of these two probabilities.

Theorem 2. Let $\Delta_i = \log(p_i/(1-p_i))$, where p_i is the prior probability that record $s_i \in S_{\text{true}}$. This choice of Δ_i minimizes the sum of the Type I and Type II error probabilities when comparing $S^* = \arg\max_S F_{\text{pen}}(S)$ and S_{true}.

The proof of Theorem 2 is in the Appendix. Next, we show that $S^* = \arg\max_S F_{\text{pen}}(S)$ may be interpreted as the MAP estimate of S_{true}.

Theorem 3. Let $\Delta_i = \log(p_i/(1-p_i))$, where p_i is the prior probability that record $s_i \in S_{\text{true}}$. This choice of Δ_i makes $S^* = \arg\max_S F_{\text{pen}}(S)$ the MAP estimate of the true affected subset S_{true}.

Proof of Theorem 3.

$$\begin{aligned}
\log \Pr(H_1(S) \mid D) &\propto \log \Pr(D \mid H_1(S)) + \log \Pr(H_1(S)) \\
&\propto F(S) + \log\left(\prod_{s_i \in S} p_i \prod_{s_i \notin S} (1-p_i)\right) \\
&= F(S) + \sum_{s_i \in S} (\log p_i - \log(1-p_i)) + \sum_{i=1}^{N} \log(1-p_i) \\
&= F(S) + \sum_{s_i \in S} \Delta_i - \sum_{i=1}^{N} \log(1+\exp(\Delta_i)) \\
&\propto F(S) + \sum_{s_i \in S} \Delta_i,
\end{aligned}$$

where terms independent of S have been ignored. Thus, choosing the subset S^* that maximizes $F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$ also maximizes the posterior probability of $H_1(S)$ making S^* the MAP estimate of S_{true}.

This Bayesian interpretation of the penalized maximum likelihood estimate should not be confused with the Bayesian and multivariate Bayesian scan statistics (Neill, Moore, and Cooper 2006; Neill and Cooper 2010), which calculate marginal likelihoods and compute the total posterior probability that each subset S has been affected. The Bayesian scan framework in previous work is limited to Gamma-Poisson count data and cannot be easily generalized to other settings.

3.2 Efficient Optimization of the Penalized Score Function

We now consider how the optimal *penalized* subset $S^* = \arg\max_{S \subseteq D} F_{\text{pen}}(S)$ can be efficiently computed. As noted above in Corollary 2, for a given value of the relative risk q, $F_{\text{pen}}(S \mid q)$ can be efficiently optimized over subsets by including all and only those data elements making a positive contribution to the penalized scoring function, that is, those

data elements with $\gamma_i(q) = \lambda_i(q) + \Delta_i > 0$. We now show that only linearly rather than exponentially many values of q must be considered.

Theorem 4. The optimal subset $S^* = \arg\max_S F_{\text{pen}}(S)$ maximizing a penalized expectation-based scan statistic from the exponential family may be found by evaluating only $O(N)$ subsets, where N is the total number of data elements.

Proof of Theorem 4. Let $\gamma_i(q) = \lambda_i(q) + \Delta_i$ as defined above, and assume that all Δ_i are independent of q. The first derivative $\gamma_i'(q) = \lambda_i'(q) = \mu_i[T(x_i) - q\mu_i]\theta'(q\mu_i)$ has only one zero, obtained when $q = T(x_i)/\mu_i$ (the maximum likelihood estimate). Thus, $\gamma_i(q)$ has at most two zeros. More precisely, we must have either (a) there exists some $q_i^{\min}$ and $q_i^{\max}$ such that $\gamma_i(q_i^{\min}) = \gamma_i(q_i^{\max}) = 0$ and $\gamma_i(q) > 0$ for all $q_i^{\min} < q < q_i^{\max}$, or (b) for all q, $\gamma_i(q) \leq 0$. In the latter case, data element s_i will never be included in the highest-scoring penalized subset. Critically, we must consider at most $2N$ distinct values of q, since we consider the $q_i^{\min}$ and $q_i^{\max}$ for each s_i, $i = 1, \ldots, N$. This property also holds when restricting $q > 1$; see Figure 1 for an example using the penalized expectation-based Poisson scan statistic. We now sort these values of q (eliminating any duplicate q values) and let $I_1, \ldots, I_{2N}$ be the disjoint intervals formed by consecutive values of the sorted q. By construction, within each interval I_j, we have for each s_i that either $\gamma_i(q) > 0$ for all $q \in I_j$, in which case including s_i will increase the penalized score for all values of q in this interval, or $\gamma_i(q) < 0$ for all $q \in I_j$, in which case including s_i will decrease the penalized score for all values of q in this interval. Also, we note that if $q \notin \bigcup_{j=1}^{2N} I_j$, then $F_{\text{pen}}(S \mid q) \leq 0$ for all S, and hence we only need to evaluate the best subset for $q \in \bigcup_{j=1}^{2N} I_j$. We can write

$$\begin{aligned} S^* &= \arg\max_S \max_{q>1} F_{\text{pen}}(S \mid q) \\ &= \arg\max_S \max_{q \in \bigcup_{j=1}^{2N} I_j} F_{\text{pen}}(S \mid q) \\ &= \arg\max_{j \in \{1,\ldots,2N\}} F_{\text{pen}}(S_j^*), \end{aligned}$$

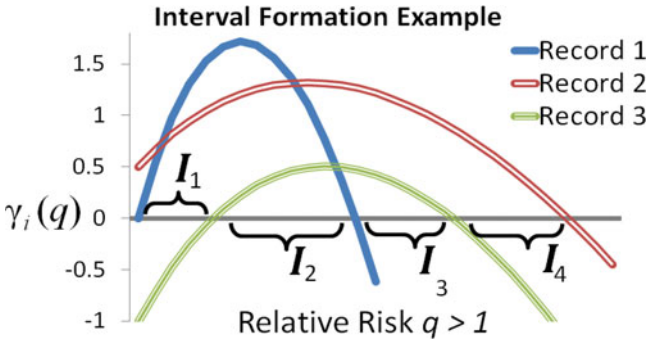

Figure 1. A three-record example of forming the $O(N)$ intervals needed to evaluate $F_{\text{pen}}(S \mid q)$. Throughout interval I_1, records 1 and 2 are making positive contributions and would be included in S_1^*. S_2^* would include all three records. S_3^* would include records 2 and 3, and S_4^* would include record 2 only. Further details: $x_1 = 130$, $\mu_1 = 110$, $\Delta_1 = 0$; $x_2 = 26$, $\mu_2 = 20$, $\Delta_2 = 0.5$; $x_3 = 40$, $\mu_3 = 30$, $\Delta_3 = -1$. $I_1 = [1, 1.132]$, $I_2 = [1.132, 1.3844]$, $I_3 = [1.3844, 1.557]$, and $I_4 = [1.557, 1.760]$.

where $S_j^* = \arg\max_S F_{\text{pen}}(S \mid q \in I_j)$. We can construct these sets efficiently as follows:

$$S_j^* = \{s_i : \gamma_i(q) > 0 \text{ for all } q \in I_j\}.$$

Note that S_j^* is the set of all elements that make positive contributions to the score $F_{\text{pen}}(S \mid q \in I_j)$ through $\gamma_i(q)$. Hence, S_j^* is an optimal subset for any $q \in I_j$. Therefore, we need to evaluate only $O(N)$ subsets (one for each interval) to find the optimal penalized subset S^*.

4. RELATIONSHIP BETWEEN ALTSS AND LTSS

The LTSS property enables exact and efficient optimization of *un*penalized scoring functions from the "separable" exponential family. Insights from the ALTSS property expand on the LTSS property in two ways. First, we consider an alternative priority function that enables us to broaden the class of functions that satisfy LTSS to expectation-based scan statistics from the entire exponential family. Second, our PFSS framework introduced in Section 3 enables exact and efficient optimization of both penalized and unpenalized score functions, while LTSS applies only in the unpenalized case.

A function satisfies LTSS if and only if $\max_{S \subseteq D} F(S) = \max_{j=1,\ldots,N} F(\{s_{(1)}, \ldots, s_{(j)}\})$ where $s_{(j)}$ represents the jth highest priority data element according to a provided priority function (Neill 2012). The highest scoring subset must be composed of the j highest priority data elements for some priority function $g(s_i)$ and some j between 1 and N. Neill (2012) defined a "separable" subfamily of the exponential family, including those distributions such as the Poisson, Gaussian, and exponential for which $\theta(q\mu_i)$ can be expressed as $z_i\theta_0(q) + v_i$, for z_i and v_i independent of q. He then proved that expectation-based scan statistics from the separable exponential family satisfy LTSS with the priority function $g(s_i) = \frac{x_i}{\mu_i}$. In other words, the highest scoring subset must consist of the j data elements s_i with largest ratios of x_i to μ_i, for some j between 1 and N. The ratio of observed counts to expected counts is also the maximum likelihood estimate of the relative risk q for the individual record s_i. This is referred to as q_i^{mle}.

The binomial distribution, while part of the exponential family, is not included in the separable exponential family. We show that the expectation-based binomial (EBB) scan statistic cannot be efficiently optimized using the priority function q_i^{mle}. EBB assumes that each count x_i is drawn from a binomial distribution $\text{Bin}(n_i, p_i)$, with mean $\mu_i = n_i p_i$, under H_0. See Table 1 for more details. Consider a dataset with three elements $\{s_1, s_2, s_3\}$, where $(x_1, \mu_1, n_1) = (1500, 300, 4000)$; $(x_2, \mu_2, n_2) = (25, 8, 40)$; and $(x_3, \mu_3, n_3) = (12, 4, 40)$. The priority function $g(s_i) = \frac{x_i}{\mu_i} = q_i^{\text{mle}}$ suggests $\{s_1\}$, $\{s_1, s_2\}$, and $\{s_1, s_2, s_3\}$ as the three subsets to evaluate. However, the subset that maximizes the EBB statistic is $S^* = \{s_1, s_3\}$, with $F(S^*) \approx 1437$ at $q = 4.97$. We note that s_2 would make a negative contribution to the score for all $q \geq 4.7$, so it is not included in S^*.

We now provide an alternative priority function that satisfies LTSS for unpenalized scoring functions from the single-parameter exponential family including the binomial and negative binomial distributions. Recall that for each record s_i there exists $q_i^{\min}$ and $q_i^{\max}$ such

that $\lambda_i(q_i^{\min}) = \lambda_i(q_i^{\max}) = 0$. For unpenalized scoring functions ($\Delta_i = 0$), $q_i^{\min} = 1$ for all i, while $q_i^{\max}$ is a function of the observed count x_i and expected count μ_i.

Theorem 5. Unpenalized expectation-based scan statistics from the single-parameter exponential family satisfy the LTSS property with priority function $g(s_i) = q_i^{\max}$, where $q_i^{\max}$ is the unique $q > 1$ such that $\lambda_i(q_i^{\max}) = 0$.

Proof of Theorem 5. We denote the jth highest priority record as $s_{(j)}$ with $s_{(1)}$ as the highest priority record and $s_{(N)}$ as the lowest. We write the priority of record $s_{(j)}$ as $g(s_{(j)}) = q_{(j)}^{\max}$. Assume that the jth priority record $s_{(j)}$ is included in the optimal subset S^*. It suffices to show that all higher priority records, $s_{(1)}, \ldots, s_{(j-1)}$, must also be included in S^*. By Theorem 1, we know that expectation-based scan statistics from the exponential family satisfy ALTSS and may be written as *additive functions* over the data elements contained in the subset for a fixed risk q. By Corollary 2, we know that if $s_{(j)} \in S^*$ then there exists a fixed relative risk q^*, where $q^* = \arg\max_{q>1} F(S \mid q)$, such that the jth highest priority record is making a positive contribution at that risk, $\lambda_{(j)}(q^*) > 0$. Furthermore, we have $q_{(j)}^{\min} = 1 < q^* < q_{(j)}^{\max}$. Finally, consider any higher priority record $s_{(h)}$ and note that the priority ordering implies $q_{(h)}^{\max} > q_{(j)}^{\max}$. It follows that $s_{(h)}$ must also have $q_{(h)}^{\min} = 1 < q^* < q_{(h)}^{\max}$, which implies $\lambda_{(h)}(q^*) > 0$ and therefore $s_{(h)} \in S^*$.

In summary, using priority function $g(s_i) = q_i^{\max}$, we have shown that inclusion of the jth highest priority record in the highest scoring subset necessitates the inclusion of all higher priority records. Therefore, the optimal subset may be efficiently identified by sorting the records based on $q_i^{\max}$ and evaluating only the N subsets of the form $\{s_{(1)}, \ldots, s_{(j)}\}$ for $j = 1, \ldots, N$. Figure 2 provides a visual comparison for the expectation-based Poisson (EBP) and binomial scoring functions and the two priority functions $q_i^{\max}$ and q_i^{mle} discussed in this section.

Theorem 5 shows a connection between LTSS and ALTSS for unpenalized scoring functions. We now provide a *penalized* scoring function that satisfies ALTSS but not LTSS. Consider maximizing the expectation-based Poisson scoring function with a penalty on subset size, $F_{\text{pen}}(S) = F_{\text{EBP}}(S) - |S|$. The unpenalized EBP scoring function satisfies the LTSS property, but including the size penalty violates LTSS, preventing the efficient optimization of the penalized scoring function over subsets of the data. Consider a dataset with three elements: $(x_1, \mu_1) = (5, 2)$, $(x_2, \mu_2) = (68, 55)$, and $(x_3, \mu_3) = (68, 55)$. Note that s_1 is the optimal penalized subset of $\{s_1, s_2\}$ so if $F_{\text{pen}}(S)$ satisfies LTSS, s_1 must be higher priority than s_2. However, the highest scoring penalized subset of $\{s_1, s_2, s_3\}$ is $\{s_2, s_3\}$. This implies that s_2 must be higher priority than s_1, which is a contradiction. Thus, no priority function can exist for which $F_{\text{pen}}(S)$ satisfies the LTSS property.

This penalized scoring function *does* satisfy ALTSS: $F(S) = \max_{q>1} \sum_{s_i \in S}(\lambda_i(q) + \Delta_i)$, where $\lambda_i(q) = x_i \log q + \mu_i(1 - q)$ from Table 1, and $\Delta_i = -1$ for all data elements s_i, since $\sum_{s_i \in S}(-1) = -|S|$. This enables us to efficiently maximize the penalized scoring function in our new PFSS framework. Due to the Δ_i penalty terms, we no longer have $q_i^{\min} = 1$ for all i. As shown in Theorem 3, this creates a *partitioning* over q instead of a *priority ordering* over q. The partitioning creates at most $2N$ intervals over the range of $q > 1$, and for each interval we need only to consider the subset of records making

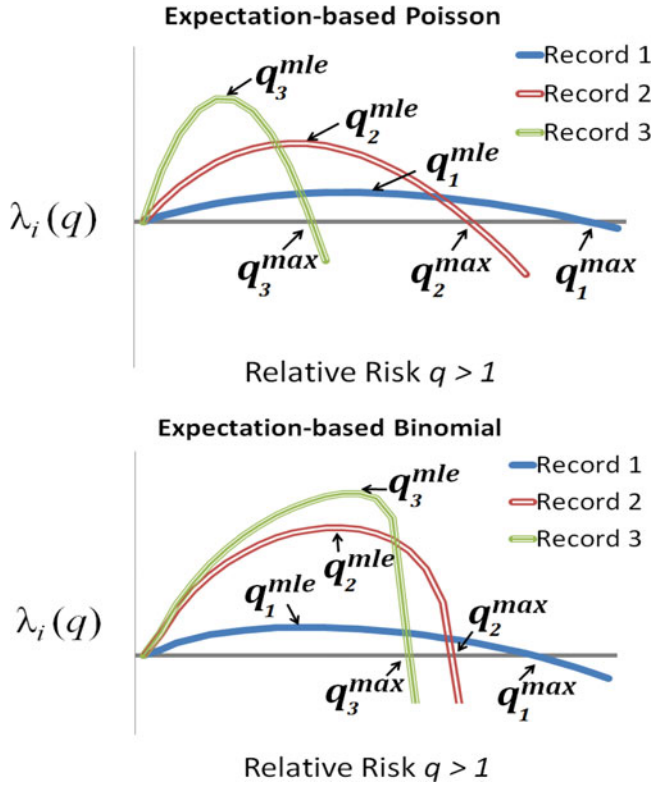

Figure 2. The top panel provides a three-record example using the expectation-based Poisson scoring function, which is a member of the "separable" exponential family. In this setting, both priority functions $g(s_i) = q_i^{\text{mle}}$ (introduced in Neill 2012) and $g(s_i) = q_i^{\text{max}}$ (introduced in this work) satisfy LTSS. Note the same ordering produced by either function. Further details: $x_1 = 8$, $\mu_1 = 6$, $q_1^{\text{mle}} = 1.33$, $q_1^{\text{max}} = 1.74$; $x_2 = 35$, $\mu_2 = 28$, $q_2^{\text{mle}} = 1.25$, $q_2^{\text{max}} = 1.54$. $x_3 = 170$, $\mu_3 = 150$, $q_3^{\text{mle}} = 1.133$, $q_3^{\text{max}} = 1.28$. The bottom panel provides a three-record example from the expectation-based binomial scoring function, which is *not* a member of the "separable" exponential family. In this setting, the two priority functions result in different orderings. We prove in Theorem 5 that $g(s_i) = q_i^{\text{max}}$ is the correct priority ordering to satisfy LTSS for expectation-based scan statistics formed by distributions from the entire exponential family. Further details: $x_1 = 40$, $n_1 = 140$, $p_1 = 0.075$, $q_1^{\text{mle}} = 3.81$, $q_1^{\text{max}} = 7.95$; $x_2 = 125$, $n_2 = 190$, $p_2 = 0.15$, $q_2^{\text{mle}} = 4.39$, $q_2^{\text{max}} = 6.51$; $x_3 = 130$, $n_3 = 155$, $p_3 = 0.18$, $q_3^{\text{mle}} = 4.66$, $q_3^{\text{max}} = 5.555$.

a positive contribution to the scoring function. These $2N$ subsets are the only ones that must be evaluated to identify the highest scoring *penalized* subset in the PFSS framework. This partitioning of q intervals rather than use of a priority function differentiates the contributions from ALTSS in this work and LTSS in previous work. Table 2 summarizes the comparison of LTSS and ALTSS.

5. PFSS WITH SOFT PROXIMITY CONSTRAINTS

The fast localized scan (Neill 2012) performs searches for each local neighborhood (center location s_c and its $k - 1$ nearest neighbors), thus enforcing hard constraints on spatial proximity.

Table 2. Summary of the LTSS and ALTSS comparisons

Scoring functions	Priority function	Number of subsets to be evaluated	Notes
Separable exponential family with no penalty terms	(a) $g(s_i) = \frac{x_i}{\mu_i}$ or (b) $g(s_i) = q_i^{\max}$	N	(a) is from Neill 2012. (b) is proposed here.
Entire exponential family with no penalty terms	$g(s_i) = q_i^{\max}$	N	We expand the class of scoring functions that satisfy LTSS.
Entire exponential family with penalty terms	No priority function satisfies LTSS	$2N$	We introduce ALTSS to efficiently incorporate penalty terms.

PFSS with soft proximity constraints allows us to take additional spatial information into account, rewarding spatial compactness and penalizing sparse regions *within* a local neighborhood. When considering a local neighborhood S_{ck} with center location s_c and neighborhood size k, we define Δ_i for each location $s_i \in S_{ck}$ as: $\Delta_i = h\,(1 - (2d_i/r))$, where d_i is the distance between location s_i and the center location s_c, r is the neighborhood radius (distance from s_c to its $(k-1)$th neighbor), and $0 \leq h \leq \infty$ is a constant representing the strength of the soft proximity constraint. Through the prior log-odds interpretation of Δ_i, we interpret h as assuming that the center location ($d_i = 0$, $\Delta_i = h$) is $\exp(h)$ times as likely to be included in the affected subset as its $(k-1)$th neighbor ($d_i = r$, $\Delta_i = -h$). Figure 3 shows the probability of inclusion for locations that are a distance d_i from the center, across various values of h. Note that for $h = 0$, PFSS reduces to the original FSS solution. Incorporation of soft proximity constraints ($h > 0$) gives preference to more spatially compact clusters by rewarding locations that are closer to the center, while still considering all subsets within a given neighborhood. For very large h values, all locations with $d_i < r/2$ would have $\Delta_i >> 0$ and all locations with $d_i > r/2$ would have $\Delta_i << 0$, and thus PFSS reduces to the circular scan with fixed radius $r/2$.

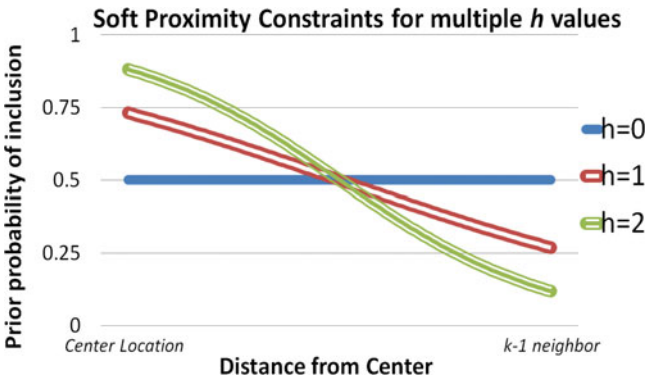

Figure 3. A location's prior probability of being included in the detected subset is based on both h and its distance from the center location. Note that the center location is assumed to be $\exp(h)$ times more likely to be included than the farthest $(k-1$ neighbor) location.

The PFSS algorithm with soft proximity constraints (Algorithm 1) builds a local neighborhood of size k for each center location, then computes the penalties Δ_i and maximizes the penalized scoring function $F_{\text{pen}}(S)$ for each neighborhood. Note that we are considering each location s_i, $i = 1, \ldots, N$, as a possible center. To compare scores across different neighborhoods, we subtract the sum $\sum_{s_i \in S_{ck}} \log(1 + \exp(\Delta_i))$. This insures that $F_{\text{pen}}(S)$ is proportional to the log-posterior probability $\Pr(H_1(S) \mid D)$, and thus we maintain the interpretation of $S^* = \arg\max_S F_{\text{pen}}(S)$ as a MAP estimate (Theorem 3).

Algorithm 1: Penalized Fast Subset Scanning with soft proximity constraints.

for $c = 1, \ldots, N$ **do**
  Let S_{ck} be center location s_c and its $k - 1$ nearest neighbors.
  for each $s_i \in S_{ck}$ **do**
    Compute Δ_i, $q_i^{\min}$, and $q_i^{\max}$.
  end for
  $Q \leftarrow$ sort and remove duplicates($\{q_1^{\min}, q_1^{\max}, \ldots, q_{2k}^{\min}, q_{2k}^{\max}\}$).
  If there exists any $q \in Q$ such that $q < 1$, exclude all $q < 1$ and add $q = 1$ to Q.
  $S \leftarrow \{\emptyset\}$.
  for $j = 1, \ldots, 2k$ **do**
    If Q_j is a $q_i^{\min}$, then $S \leftarrow S \cup \{s_i\}$. If Q_j is a $q_i^{\max}$, then $S \leftarrow S \setminus \{s_i\}$.
    Record $F_{\text{pen}}(S) = F(S) + \sum_{s_i \in S} \Delta_i$.
  end for
  Subtract $\sum_{s_i \in S_{ck}} \log(1 + \exp(\Delta_i))$ from $F_{\text{pen}}(S)$.
end for
Output the optimal subset $S^* = \arg\max_S F_{\text{pen}}(S)$.

We conclude this section with a complexity analysis for PFSS. To find the optimal subset for a given neighborhood, we sort the at most $2k$ values of q, which is an $O(k \log k)$ operation, and step through the sorted values of q, which is an $O(k)$ operation. Over N neighborhoods, the total computational complexity of this algorithm is $O(Nk \log k)$. This assumes that the k-nearest neighbors have been precomputed for each location, since this is a one-time operation; otherwise, computation of the k-nearest neighbors of each location can be done naively in $O(N^2 \log N)$ or more quickly using space-partitioning data structures. PFSS was able to identify the highest scoring penalized subset for a single day of our Emergency Department data described in Section 7 (with $N = 97$ locations) in 40–50 milliseconds for all values of $k = 5, \ldots, 50$, which is comparable to the runtimes of the original FSS and the circular spatial scan.

6. RELATED WORK

PFSS with soft proximity constraints combines penalized likelihood ratio statistics, spatial data, and subset scanning to increase detection power for irregularly shaped spatial

clusters. The subset scanning approach is unique in separating this present work from methods that also use spatial information and attempt to optimize a penalized likelihood ratio statistic. For example, Yiannakoulias, Rosychuk, and Hodgson (2005) penalized nonconnected search regions, while Duczmal et al. (2007) and Kulldorff et al. (2006) computed the geometric regularity of the search region and penalized more elongated and irregularly shaped clusters. More sophisticated methods combine geometric and nonconnectivity penalties in a multi-objective framework (Cancado et al. 2010). However, most of these methods rely on a heuristic search to optimize the penalized scan statistic, which is computationally expensive and not guaranteed to identify the highest-scoring cluster, while Kulldorff et al. (2006) limited their search to elliptical clusters, reducing detection power, and spatial accuracy for any subsets that are not well approximated by an ellipse. In contrast, our PFSS approach is extremely computationally efficient and scalable while guaranteeing that the highest-scoring penalized subset will be found. It is also worth noting that the previously proposed methods focus on penalizing or rewarding properties of the region as a whole rather than individual data elements, while our penalties at the data-element level have a direct interpretation as the prior log-odds for each element's inclusion in the optimal subset. Either of these types of penalty could be preferable for a given application domain.

We note that the ALTSS property is distinct from prior work in submodular function optimization (Nemhauser, Wolsey, and Fisher 1978; Leskovec et al. 2007), which has been used for sensor placement among many other applications. As shown by Neill (2012), the expectation-based Poisson statistic does not satisfy submodularity. Further, methods based on submodularity typically find approximate rather than exact solutions, while our approach is guaranteed to find the optimal subset that maximizes the penalized statistic.

7. EVALUATION

We provide an example for the PFSS method with soft proximity constraints in the public health surveillance domain. Emergency Department data from 10 Allegheny County, Pennsylvania hospitals from January 1, 2004, to December 31, 2005, serves as the background data for both validation of the PFSS framework and a performance evaluation for detecting aerosolized anthrax bio-attacks. Through processing of each case's International Classification of Diseases (ICD-9) code and the free text in its "chief complaint" string, a count dataset was created recording the number of patient records with respiratory symptoms (such as cough or shortness of breath) for each day and each zip code in Allegheny County. The dataset had a daily mean of 44.0 cases, and a standard deviation of 12.1 cases. The latitude and longitude coordinates of the centroid of the $N = 97$ zip codes formed the spatial component of the dataset.

To validate the PFSS approach, we examine a simulation that varies the size and spatial density of the affected region (i.e., subset of zip codes with additional counts injected into the background data) and thus understand the effects of these parameters on the relative performance of the competing methods. We then evaluate the detection performance of PFSS using state-of-the-art dispersion models of an aerosolized anthrax release (Hogan et al. 2007). Both experiments and their results are discussed in their respective sections

below. We use the expectation-based Poisson likelihood ratio statistic throughout, and compare three methods in each setting:

- Kulldorff's circular spatial scan statistic (circles), which returns the highest scoring circular region, searching over all N distinct circles with neighborhood size k centered at the N locations (Kulldorff 1997).
- Fast subset scan (FSS) that returns the highest scoring *un*penalized subset within a region consisting of a center location and its $k-1$ nearest neighbors for a fixed parameter k (Neill 2012). This can be considered a special case of PFSS with the strength of the soft proximity constraint $h = 0$.
- Penalized fast subset scan (PFSS) with soft proximity constraints, which returns the highest scoring *penalized* subset within a region consisting of a center location and its $k-1$ nearest neighbors for a fixed parameter k. The soft proximity constraints reward spatial compactness while penalizing sparse regions. We provide results for both weaker ($h = 1$) and stronger ($h = 2$) constraints. Choice of h is discussed below.

We consider two evaluation metrics: detection power (proportion of outbreaks detected) at a fixed false positive rate of 1 per year and spatial accuracy measured by the "overlap coefficient" between true and detected clusters. Overlap is a combination of precision and recall and requires two sets, S_{true} of affected locations and S^* of detected locations. Then the overlap coefficient is defined as: Overlap $= |S_{\text{true}} \bigcap S^*|/|S_{\text{true}} \bigcup S^*|$. An overlap coefficient of 1 (or 100%) represents perfect precision and recall, while an overlap of 0 corresponds to disjoint sets S_{true} and S^*.

For the three methods considered here, Type I error can be controlled by the use of randomization testing or empirical calibration to set the threshold score for detection. For example, for a fixed false positive rate of 1/year, we would set the threshold score at the $100(1 - \frac{1}{365}) \approx 99.7$th percentile of the distribution of daily maximum scores under the null. We can then compare the methods' detection power for the given Type I error rate.

7.1 VALIDATION ON SIMULATED OUTBREAKS

We create a large set of simple simulated outbreaks for validation to compare the relative performance of PFSS, FSS, and the circular spatial scan (circles) as a function of outbreak size, spatial density, and neighborhood size k. For each simulated outbreak, the simulator selects the affected subset of zip codes S_{true} uniformly at random (between 5 and 10 affected zip codes). Then Poisson($w_i\ |S_{\text{true}}|$) additional cases are injected into each location in S_{true}, where $w_i = c_i/(\sum_{s_j \in S_{\text{true}}} c_j)$ represents the relative "weight" of zip code s_i, proportional to the total number of cases in that zip code for the entire 2 years of Emergency Department data. The simulated outbreaks are categorized by spatial density, measured by the ratio of the number of affected locations to the total number of locations in the smallest circle that contains all affected locations and size measured by the total number of affected locations. Figure 4 provides two examples. Results for nine scenarios are provided; three categories of density (0.1–0.4 for "low," 0.4–0.7 for "medium," and 0.7–1.0 for "high") and three categories of size based on the number of affected zip codes (5–6 for "small," 7–8 for "medium," and 9–10 for "large").

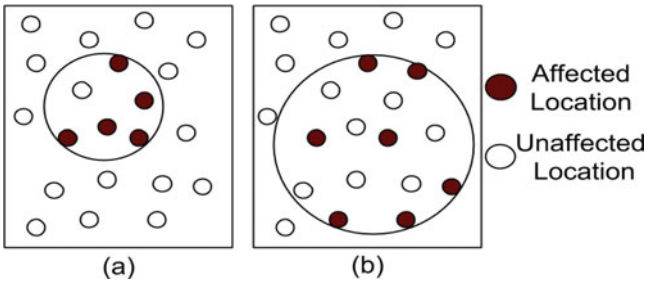

Figure 4. Two examples of inject density. Figure (a) is a region with a density of 5/6 and (b) is a region with density 7/14.

Figures 5 and 6 have the same layout with spatial density increasing between panels from left to right and outbreak size increasing between panels from bottom to top. The lower left panel of few, highly dispersed affected zip codes represents the most difficult detection scenario while the upper right panel of many, highly compact affected zip codes reflects the easiest scenario.

Figure 5 provides a comparison of detection power (proportion of outbreaks detected at 1 false positive per year). As expected, the overall performance for all methods increases with the number and spatial density of the affected zip codes. We note the poor performance of the spatial scan statistic (circles) for the low density outbreaks. This is due to only scanning over circular regions, which results in much lower detection power for irregularly shaped clusters. The spatial scan statistic performs comparably in the high density outbreaks that are compact and close to circular in shape. We also examine the effect that the neighborhood size k has on the methods and note that the detection power of FSS is heavily influenced by the choice of k. The influence of k is more pronounced in outbreaks composed of few, compact affected zip codes (lower right panel). In contrast, we note that the detection power of PFSS remains strong for a wide range of neighborhood sizes, densities, and numbers of affected locations. Despite the lack of spatial structure in the low density outbreaks, the

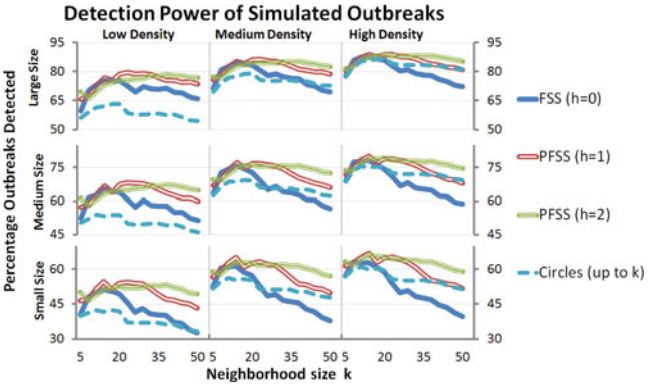

Figure 5. Comparison of detection power for multiple methods at a fixed false positive rate of 1 per year. Each panel represents different outbreak spatial density and size. Neighborhood sizes from $k = 5, \ldots, 50$ are provided within each panel.

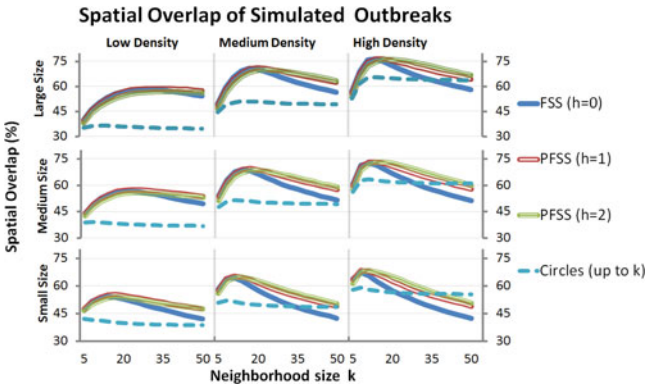

Figure 6. Comparison of spatial accuracy (overlap coefficient) for multiple methods. Each panel represents different outbreak spatial density and size. Neighborhood sizes from $k = 5, \ldots, 50$ are provided within each panel.

penalized methods (which reward spatially compact subsets) outperform the unpenalized method, FSS. We attribute this strong performance to PFSS's robustness to noise in the background data, increasing overall detection power. For large values of k, FSS is more likely to give high scores to spatially dispersed subsets in the background data, increasing the threshold needed to detect the simulated events, while PFSS will only identify such spurious clusters if they happen to be spatially localized.

Figure 6 provides a comparison of spatial accuracy. We note that larger, more spatially compact outbreaks result in higher spatial accuracy for all methods. The circular spatial scan statistic consistently underperforms FSS and PFSS, particularly for the low density clusters. It tended to return overly large circular regions with high recall but low precision, resulting in a low overlap coefficient. The robustness of the PFSS methods is shown again for the low density outbreaks. Although low density injects have a relative lack of the spatial structure that PFSS is designed to reward, the ability of PFSS to penalize sparse regions increases spatial precision while maintaining reasonably high recall, resulting in spatial accuracy that is comparable to FSS.

7.2 EVALUATION ON BARD ANTHRAX ATTACKS

The anthrax attacks are based on a state-of-the-art, highly realistic simulation of an aerosolized anthrax release, the Bayesian Aerosol Release Detector (BARD) simulator (Hogan et al. 2007). These complex simulations take into account weather data when creating the affected zip codes, S_{true}, and demographic information when calculating the number of additional Emergency Department cases within each affected zip code. Wind direction, wind speed, and atmospheric stability all influence the elongated shape and size of the affected area. Although the simulator produces data for a 10-day period after the spores are released, we simplify the temporal component by using only the data from the midpoint (day 5) of the simulation.

We consider two coverage scenarios. In the 100% coverage case, we assume that all of the anthrax victims present at an Emergency Department with a functioning bio-surveillance

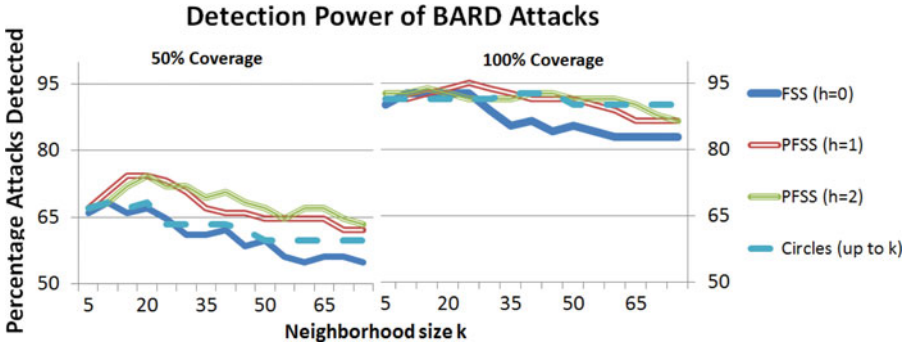

Figure 7. Comparison of detection power for multiple methods on simulated anthrax bio-attacks at a fixed false positive rate of 1 per year, for 50% and 100% coverage scenarios, respectively. Neighborhood sizes from $k = 5, \ldots, 75$ are provided within each panel.

program and are appropriately accounted for. This assumption is extremely optimistic, so we provide a possibly more realistic 50% coverage case where half of the population of anthrax victims seek medical attention from institutions that do not collect or share this type of data, creating a more difficult detection problem.

Figure 7 provides a comparison of detection power for anthrax attacks. Intuitively, the optimistic 100% coverage scenario has higher detection rates for all methods. In the more difficult 50% coverage setting, the penalized scoring functions show higher detection rates and greater robustness to the choice of neighborhood size parameter, k. The unpenalized FSS method struggles for improperly chosen k even in the easier 100% coverage scenario.

Figure 8 provides a comparison of spatial accuracy for anthrax attacks. The strong performance of the subset scanning methods compared to the circular spatial scan is due to the elongated, noncircular regions (based on assumed, randomly generated wind direction and speed) of affected zip codes produced by the BARD simulation. The performance of "circles" is similar in the 100% and 50% coverage scenarios, suggesting that it is limited by the geometry of the circular spatial scan.

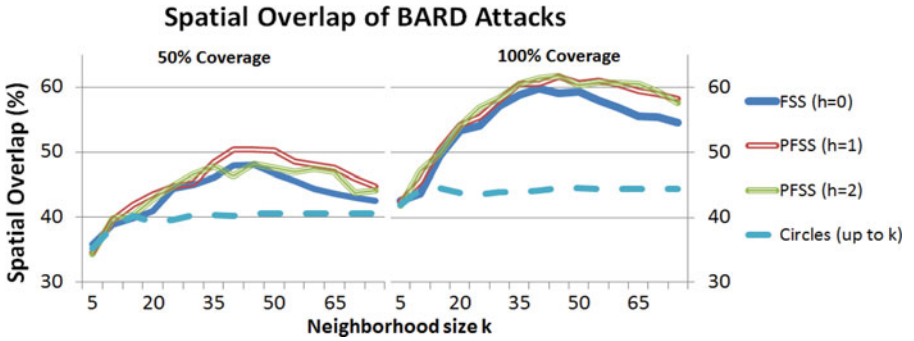

Figure 8. Comparison of spatial accuracy (overlap coefficient) for multiple methods on simulated anthrax bio-attacks, for 50% and 100% coverage scenarios, respectively. Neighborhood sizes from $k = 5, \ldots, 75$ are provided within each panel.

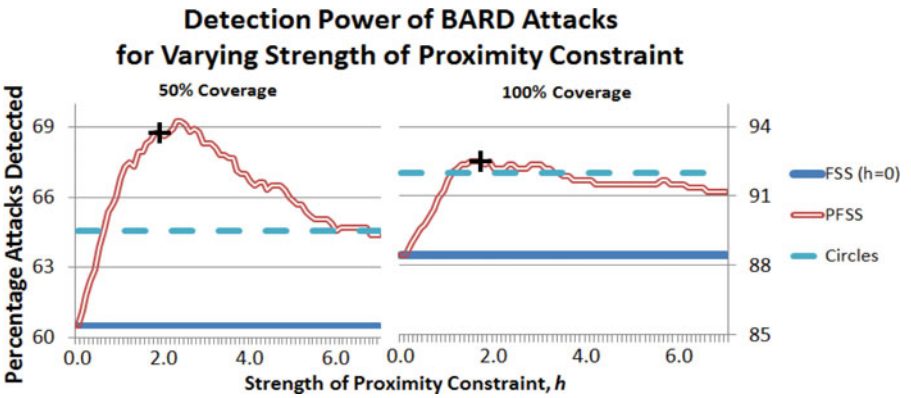

Figure 9. Comparison of detection power (averaged over all neighborhood sizes) for multiple methods on simulated bio-attacks, for 50% and 100% coverage scenarios. Soft proximity constraint strengths from $h = 0, \ldots, 7$ are provided within each panel. The black marker represents the h that maximized average detection power for a separate training dataset. Note different y-axis scales.

Figure 9 demonstrates PFSS's robustness to the choice of the proximity constraint strength, h, by comparing average detection power of the anthrax bio-attacks (averaged over neighborhood sizes $k = 5, 10, \ldots, 75$) for varying $h = 0, \ldots, 7$. For this analysis, the 82 BARD-simulated anthrax attacks were split into separate training and test groups. The black cross represents the value of h that maximized average detection power for the training dataset.

In both coverage scenarios, we note the strong performance of PFSS as compared to FSS for all values of $h = 0, \ldots, 7$. This increased performance is a combination of the robustness of PFSS to choice of h and the sensitivity of FSS to poorly chosen neighborhood size, k. The circular scan is also robust to neighborhood size k and therefore performs comparably to PFSS in the 100% coverage scenario. We note that near-optimal values of h can be learned from a small number of labeled training examples. The learned $h = 1.8$ and $h = 1.7$, for 50% and 100% scenarios, respectively, out-performed circles and FSS when evaluated on held out test data.

8. CONCLUSIONS

This work introduced and formalized the additive linear-time subset scanning (ALTSS) property, which allows exact and efficient optimization of penalized likelihood ratio scan statistics over all subsets of data elements. We demonstrated that this property holds for expectation-based scan statistics from the exponential family, thus providing flexibility to incorporate other parametric scan statistic models such as the Gaussian, exponential, binomial, and negative binomial scans. We note that this result is more general than Neill (2012) in two aspects: the incorporation of penalty terms and the extension to the entire exponential family (rather than the "separable" exponential family defined by Neill).

We incorporated this property into a penalized fast subset scan (PFSS) framework, which enables the scan statistics to be efficiently optimized including additional, element-specific penalty terms. Our critical insight is that the scoring function $F(S)$ may be written as an

additive function, summing over all data elements $s_i \in S$, when conditioning on the relative risk q. This form provides two advantages. First, additional terms may be added to the statistic to represent the prior log-odds of each data element being included, while maintaining the additive structure of the scoring function. Second, optimization of either the unpenalized scan statistic $F(S \mid q)$ or the penalized scan statistic $F_{\text{pen}}(S \mid q)$ over subsets can be performed very efficiently, by including all and only those records making a positive contribution to the score. Moreover, we demonstrated that only a small (linear rather than exponential) number of values of the relative risk q must be considered, making the computation of the highest scoring penalized subset $S^* = \arg\max_S \max_{q>1} F_{\text{pen}}(S \mid q)$ computationally tractable. Unpenalized likelihood ratio statistics from the exponential family can be optimized while only considering N subsets; penalized likelihood ratio statistics can be optimized while considering at most $2N$ subsets, and finding the exact solution is guaranteed in both cases. If the alternative hypothesis $H_1(S)$ is true for some subset S, the highest scoring penalized subset can be interpreted as a MAP estimate of the true affected subset.

As a straightforward application of our PFSS framework, we developed "soft" constraints on spatial proximity (i.e., for a given local neighborhood under consideration, locations closer to the center are assumed to be more likely to have been affected). We then applied PFSS with soft proximity constraints to the task of detecting anthrax bio-attacks, comparing its detection power and spatial accuracy to the current state of the art. PFSS demonstrated strong results, outperforming the traditional, circular spatial scan statistic (Kulldorff 1997), and the FSS (Neill 2012) in both detection power and spatial accuracy. Compared to fast subset scan (FSS), PFSS showed remarkable robustness to selection of the neighborhood size k, and this robustness extended even to low density outbreaks designed to challenge the use of soft proximity constraints.

Our PFSS framework with soft constraints introduced a parameter h for the strength of the spatial proximity constraint. The extreme cases of $h = 0$ and $h \to \infty$ correspond to the unpenalized FSS and a fixed-radius circular scan, respectively. In this work, we showed that near-optimal values of h can be learned from a small number of labeled training examples ($\sim$40). Additionally, PFSS demonstrated robustness to the choice of h, outperforming FSS for all values $h = 0, \ldots, 7$.

Soft proximity constraints serve as one example of many different applications that can take advantage of including additional prior information in the subset scanning framework. Another is to incorporate "temporal consistency constraints," which use prior information based on records that were included in the highest scoring subset at a previous time step, to increase detection power for *dynamic* patterns where the affected subset changes over time. Incorporating these constraints that penalize abrupt, unrealistic changes can be applied to detecting a contaminant spreading in a water distribution system (Speakman, Zhang, and Neill 2013).

APPENDIX: MINIMIZING ERROR WITH Δ_i

We show that if we can correctly estimate the prior probability p_i for location s_i to be in the affected subset S_{true}, then setting $\Delta_i = \log(\frac{p_i}{1-p_i})$ (the prior log-odds) minimizes the total probability

of error, including both Type I errors (including location s_i in the detected subset S^* when $s_i \notin S_{\text{true}}$) and Type II errors (failing to include a location $s_i \in S_{\text{true}}$ in the detected subset S^*).

Assume $s_i \in S_{\text{true}}$ with probability p_i, and that we observe $x_i \sim \text{Dist}_1$ if $s_i \in S_{\text{true}}$ and $x_i \sim \text{Dist}_0$ if $s_i \notin S_{\text{true}}$. Moreover, assume that λ_i and Δ_i are the log-likelihood ratio and penalty for location s_i, respectively, where $\lambda_i = \log(p(x_i \mid \text{Dist}_1)/p(x_i \mid \text{Dist}_0))$ and Δ_i can be any real number. The (unconstrained) penalized subset scan will include s_i in the detected subset S^* if and only if $\lambda_i + \Delta_i > 0$. We now show that the total probability of error is minimized for $\Delta_i = \log\left(\frac{p_i}{1-p_i}\right)$:

$$\begin{aligned}
\Pr(\text{error} \mid \Delta_i) &= \Pr(s_i \in S^* \mid s_i \notin S_{\text{true}}, \Delta_i)\Pr(s_i \notin S_{\text{true}}) \\
&\quad + \Pr(s_i \notin S^* \mid s_i \in S_{\text{true}}, \Delta_i)\Pr(s_i \in S_{\text{true}}) \\
&= (1 - p_i)\Pr((\lambda_i + \Delta_i > 0) \mid s_i \notin S_{\text{true}}) + p_i \Pr((\lambda_i + \Delta_i < 0) \mid s_i \in S_{\text{true}}) \\
&= (1 - p_i)(1 - \text{CDF}_0(-\Delta_i)) + p_i \text{CDF}_1(-\Delta_i),
\end{aligned}$$

where the cumulative density functions CDF_0 and CDF_1 are defined as follows:

$$\text{CDF}_j(z) = \int_{-\infty}^{z} p(\lambda_i = k \mid x_i \sim \text{Dist}_j)\, dk, \;\; j \in \{0, 1\},$$

and we also define the corresponding probability density functions PDF_0 and PDF_1:

$$\text{PDF}_j(z) = p(\lambda_i = z \mid x_i \sim \text{Dist}_j), \;\; j \in \{0, 1\}.$$

Furthermore, we note the key property $\text{PDF}_1(z)/\text{PDF}_0(z) = \exp(z)$, since $\text{PDF}_1(z)$ and $\text{PDF}_0(z)$ are, respectively, sums of $p(x_i)$ for all x_i with corresponding $\lambda_i = z$, and for each such x_i, we know that $p(x_i \mid \text{Dist}_1)/p(x_i \mid \text{Dist}_0) = \exp(z)$.

We proceed by setting the first derivative of Pr(error) equal to 0:

$$\begin{aligned}
\frac{d\Pr(\text{error})}{d\Delta_i} &= (1 - p_i)\text{PDF}_0(-\Delta_i) - p_i \text{PDF}_1(-\Delta_i) \\
&= (1 - p_i - p_i \exp(-\Delta_i))\text{PDF}_0(-\Delta_i) = 0.
\end{aligned}$$

This expression has a single zero at $\Delta_i = \log(p_i/(1 - p_i))$. The second derivative at this point is

$$\begin{aligned}
&-(1 - p_i - p_i \exp(-\Delta_i))\, d\text{PDF}_0(-\Delta_i) + p_i \exp(-\Delta_i)\text{PDF}_0(-\Delta_i) \\
&= \left(\frac{1}{1 + \exp(\Delta_i)}\right)\text{PDF}_0(-\Delta_i) \geq 0,
\end{aligned}$$

so this is the value of Δ_i that minimizes the probability of error.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation, grants IIS - 0916345, IIS - 0911032, and IIS-0953330. Additionally, Edward McFowland III was supported by an NSF Graduate Research Fellowship (NSF GRFP-0946825) and an AT&T Labs Fellowship.

[Received October 2014. Revised February 2015.]

REFERENCES

Agarwal, D., McGregor, A., Phillips, J. M., Venkatasubramanian, S., and Zhu, Z. (2006), "Spatial Scan Statistics: Approximations and Performance Study," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 24–33. [383]

Burkom, H. S., Murphy, S. P., and Shmueli, G. (2007), "Automated Time Series Forecasting for Biosurveillance," *Statistics in Medicine*, 26, 4202–4218. [385]

Cancado, A. L., Duarte, A. R., Duczmal, L. H., Ferreira, S. J., Fonseca, C. M., and Gontijo, E. C. (2010), "Penalized Likelihood and Multi-objective Spatial Scans for the Detection and Inference of Irregular Clusters," *International Journal of Health Geographics*, 9, 55. [395]

Duczmal, L., and Assuncao, R. (2004), "A Simulated Annealing Strategy for the Detection of Arbitrary Shaped Spatial Clusters," *Computational Statistics and Data Analysis*, 45, 269–286. [383]

Duczmal, L., Cancado, A. L., Takahashi, R. H., and Bessegato, L. F. (2007), "A Genetic Algorithm for Irregularly Shaped Spatial Scan Statistics," *Computational Statistics and Data Analysis*, 52, 43–52. [395]

Hogan, W. R., Cooper, G. F., Wallstrom, G. L., Wagner, M. M., and Depinay, J.-M. (2007), "The Bayesian Aerosol Release Detector," *Statistics in Medicine*, 26, 5225–5252. [395,398]

Kulldorff, M. (1997), "A Spatial Scan Statistic," *Communications in Statistics: Theory and Methods*, 26, 1481–1496. [383,384,385,396,401]

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. (2006), "An Elliptic Spatial Scan Statistic," *Statistics in Medicine*, 25, 3929–3943. [395]

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007), "Cost-Effective Outbreak Detection in Networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 420–429. [395]

Neill, D. B. (2009a), "An Empirical Comparison of Spatial Scan Statistics for Outbreak Detection," *International Journal of Health Geographics*, 8, 20. [385]

——— (2009b), "Expectation-Based Scan Statistics for Monitoring Spatial Time Series Data," *International Journal of Forecasting*, 25, 498–517. [383]

——— (2012), "Fast Subset Scan for Spatial Pattern Detection," *Journal of the Royal Statistical Society*, Series B, 74, 337–360. [383,384,386,387,390,392,395,396,400,401]

Neill, D. B., and Cooper, G. F. (2010), "A Multivariate Bayesian Scan Statistic for Early Event Detection and Characterization," *Machine Learning*, 79, 261–282. [388]

Neill, D. B., and Moore, A. W. (2004), "Rapid Detection of Significant Spatial Clusters," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 256–265. [383]

Neill, D. B., Moore, A. W., and Cooper, G. F. (2006), "A Bayesian Spatial Scan Statistic," in *Advances in Neural Information Processing Systems*, eds. Y. Weiss, B. Schlkopf, and J.C. Platt, Vol. 18, Cambridge, MA: MIT Press, pp. 1003–1010. [388]

Neill, D. B., Moore, A. W., Sabhnani, M. R., and Daniel, K. (2005), "Detection of Emerging Space-Time Clusters," in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 218–227. [383,384]

Nemhauser, G., Wolsey, L., and Fisher, M. (1978), "An Analysis of the Approximations for Maximizing Submodular Set Functions," *Mathematical Programming*, 14, 265–294. [395]

Speakman, S., McFowland, E., and Neill, D. B. (2015), "Scalable Detection of Anomalous Patterns With Connectivity Constraints," *Journal of Computational and Graphical Statistics*, 24, 1014–1033. DOI:10.1080/10618600.2014.960926. [383]

Speakman, S., Zhang, Y., and Neill, D. B. (2013), "Dynamic Pattern Detection With Temporal Consistency and Connectivity Constraints," in *Proceedings of the 13th IEEE International Conference on Data Mining*, pp. 697–706. [401]

Wu, M., Song, X., Jermaine, C., Ranka, S., and Gums, J. (2009), "A LRT Framework for Fast Spatial Anomaly Detection," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 887–896. [383]

Yiannakoulias, N., Rosychuk, R. J., and Hodgson, J. (2005), "Adaptations for Finding Irregularly Shaped Disease Clusters," *International Journal of Health Geographics*, 6, 28. [395]