Connected Sub-graph Detection

Jing Qian
Boston University

Venkatesh Saligrama
Boston University

Yuting Chen
Boston University

Abstract

We characterize the family of connected sub-graphs in terms of linear matrix inequalities (LMI) with additional integrality constraints. We then show that convex relaxations of the integral LMI lead to parameterization of all weighted connected subgraphs. These developments allow for optimizing arbitrary graph functionals under connectivity constraints. For concreteness we consider the connected sub-graph detection problem that arises in a number of applications including network intrusion, disease outbreaks, and video surveillance. In these applications feature vectors are associated with nodes and edges of a graph. The problem is to decide whether or not the null hypothesis is true based on the measured features. For simplicity we consider the elevated mean problem wherein feature values at various nodes are distributed IID under the null hypothesis. The non-null (positive) hypothesis is distinguished from the null hypothesis by the fact that feature values on some unknown connected sub-graph has elevated mean.

1 Introduction

We consider a connected graph $G = (V, E)$ where nodes $v \in V$ are associated with features values x_v that follow some statistical distribution. Our goal is to optimize some objective function of the feature values over all connected sub-graphs. To motivate this scenario consider the disease outbreak problem [1] depicted in Fig.1. Here a cholera outbreak along a winding river can lead to elevated numbers for cases in those counties near the river, which form a connected cluster in the graph representing geographical counties. Moreover, this region can be irregularly shaped as seen in Fig.1. Many other problems arising in network intrusion and video surveillance share similar features.

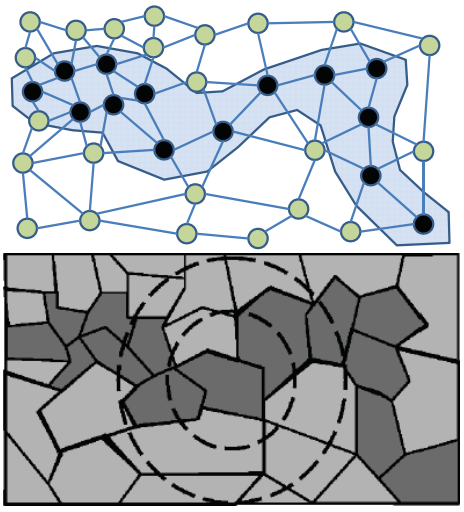


Figure 1: Graph representation of perhaps a cholera outbreak along a winding river floodplain, where each cell represents a county. The corresponding region forms a connected and irregularly shaped cluster. In the upper panel, nodes represent counties, with adjacent counties linked by edges. Dotted circles depict conventional scanning methods [2]. The figure on the lower panel is from [1].

This problem is known to be difficult [1, 3], because there does not exist systematic ways of characterizing the family of connected sub-graphs on a general graph. Existing approaches deal with this issue by optimizing some cost function over a sub-class of well-defined connected sets. For instance, scanning methods that optimize over rectangles, circles or neighborhood balls [2, 4] across different regions of the graphs are often employed. However, it has been recognized that this can result in loss of detection power [1].

The main contribution of this paper is to characterize the family of connected sub-graphs in terms of linear matrix inequalities. We present a convex parameterization of all weighted connected sub-graphs. In addition when integral constraints on the variables are im-

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

posed we show that that they result in a parameterization of all unweighted connected sub-graphs. These developments allow for optimizing arbitrary graph functionals under connectivity constraints. For concreteness we consider the connected subgraph detection problem that has recently been the subject of extensive study ([5, 6, 7, 3, 8, 9]). The problem is to decide whether or not the null hypothesis is true based on the measured features. For simplicity we consider the problem wherein feature values across the nodes are distributed IID under the null hypothesis. The non-null (positive) hypothesis is distinguished from the null hypothesis by the fact that feature values on some unknown connected sub-graph follows a different distribution. Much of this literature focuses on statistical decision aspects of the problem and usually only consider relatively simple graphs such as lines, lattices or trees, and optimize over relatively simple sub-graphs such as rectangles, balls or some other low-dimensional parametric shapes. In [10, 2, 11] general graphs are considered yet the connected shapes are still relatively simple. To deal with more complex shapes [12, 13] suggest a heuristic simulated annealing algorithm but this method requires multiple starts and often needs many iterations to achieve the global optimum. Recently, [14], has proposed a spectral scan statistic method which is based on graph regularization. While this method does not necessarily restrict shapes of connected sub-graphs, it has other disadvantages. Sub-graphs are obtained by virtue of graph partitioning. The graph partitions are not necessarily connected. Furthermore, graph regularization favors balanced size partitions and partitions with small conductance. In contrast our method guarantees connectivity and allows for arbitrarily shaped connected sub-graphs.

The paper is organized as follows. In Sec.2 we introduce the problem setup and list some examples of detection objectives that fit into our setup. We then focus on the problem of anomalous cluster detection under Gaussian model in Sec.2.1. In Sec.3 we characterize the sufficient and necessary condition for exact connectivity of a sub-graph in terms of an integral LMI constraint. We then relax it into an SDP constraint and show some nice properties of it. In Sec.4 we discuss several alternative approaches. Experiments on synthetic and real data sets are reported in Sec.5.

2 Problem Setup

Let $G = (V, E)$ denote an undirected unweighted graph with $|V| = n$ nodes and $|E| = m$ edges. Let $S \subseteq V$ be some subset of nodes. Define the indicator vector of S: $f = f^{(S)} : V \to \{0,1\}^n$, $f_i = 1$ for $i \in S$ and 0 otherwise. Let Λ be the collection of all sub-graphs of G that are connected: $\Lambda = \{S \subseteq V : S\, is\, connected\}$.

Let A denote the adjacency matrix of G. D is the degree matrix $D = diag(d)$ where d is the degree vector, $d_i = \sum_j A_{ij}$. $L = D - A$ is the unnormalized graph Laplacian matrix of G. $G_S = (S, E_S)$ denotes the induced sub-graph of G on S, with A_S, D_S, L_S the corresponding adjacency, degree and Laplacian matrices of G_S. $\mathbf{1}_n$ denotes n-dim all-one vector. $diag(x)$ is the diagonal matrix with diagonal entries equal to x; $diag(A)$ denotes the vector of diagonal entries of matrix A. $A \circ B$ denotes element-wise matrix multiplication: $(A \circ B)_{ij} = A_{ij}B_{ij}$.

We are concerned with the following problem of optimizing some cost function $c(\cdot)$ on connected sub-graphs $S \in \Lambda$ in G:

$$\max_{S \in \Lambda} : \; c(S) = c(f^{(S)}). \tag{1}$$

Notice that generally this problem is hard to solve. In fact, the prize-collecting Steiner Tree problem with equal edge cost can be reduced to Eq.(1), and is known to be NP-hard [15].

We present some examples to motivate this setup.

(1) Positive Elevated Mean Scan Statistic: The scan statistic for a connected cluster S is:

$$\eta(S) = \frac{1}{\sqrt{|S|}} \sum_{i \in S} x_i \;\; \Rightarrow \;\; c(f) = \frac{f'x}{\sqrt{f'\mathbf{1}_n}} \tag{2}$$

where x_i is the observation at node i. This statistic corresponds to the generalized likelihood ratio test (GLRT) in many contexts for 1-parameter exponential family models. The Gaussian case has a simple signal interpretation. We associate random variables with each node as follows:

$$x_i = \mu \cdot 1_{\{i \in S\}} + \epsilon_i, \; i \in V \tag{3}$$

where $\mu > 0$ is the signal strength. ϵ_i is i.i.d. standard Gaussian across different nodes. S is some unknown anomalous cluster which forms a connected component, $S \in \Lambda$. The aim is to decide between the null hypothesis $H_0 : x_i \sim \mathbb{N}(0,1), \forall i \in V$ and the alternative $H_1 = \bigcup_{S \in \Lambda} H_{1,S}$, where $H_{1,S} : x_i \sim \mathbb{N}(\mu, 1), \forall i \in S; x_i \sim \mathbb{N}(0,1), \forall i \notin S$. [3] has shown that under some conditions the test of rejecting H_0 for large values of $\max_{S \in \Lambda} : \eta(S)$ is statistically optimal or near-optimal in the minimax sense. The issue here is the computation of Eq.(1) when $c(S) = \eta(S)$.

(2) Elevated Mean Scan Statistic: When μ of Eq.(3) can be either positive or negative, the corresponding scan statistic (GLRT) can be shown to be:

$$\eta_2(S) \;\; = \;\; |\eta(S)| = \frac{1}{\sqrt{|S|}} |\sum_{i \in S} x_i| \tag{4}$$

$$\Rightarrow c(f) \;\; = \;\; \frac{f'(xx')f}{f'\mathbf{1}_n} = \frac{(xx') \circ M}{trace(M)} \tag{5}$$

where $M = ff'$.

(3) General graph functionals: The techniques developed here generalize to objectives that involve both node and edge features on connected sub-graphs S. We can incorporate this by means of a node indicator vector f, and generalizing it to edge indicator by using quadratic variants such as: $f_i f_j = 1$ for $i \in S$, $j \in S$, $(i,j) \in E$, and 0 otherwise.

2.1 Optimizing Scan Statistic for Positively Elevated Mean

We need to solve: $\max_{S\in\Lambda} \eta(S)$ of Eq.(2). Note that $\eta(S) = \eta(f)$ is not even concave in binary variables f. We propose to convexify the objective by transforming it into a 2-step procedure. This procedure involves first solving a family of sub-problems parameterized in size of S, followed by a model selection step. Convex relaxation on f will be presented in Sec.3.

Algorithm 1: Scan Statistic Computation
Input: observations $\{x_1, \ldots, x_n\}$ associated with n nodes, adjacency matrix A, size parameter set $\mathbb{K}$.

1. For different values of $k \in \mathbb{K}$, solve the following sub-routine conditioned on the size:
$$\max_{S\in\Lambda} : \sum_{i\in S} x_i = f'x \quad s.t.\ |S| = f'\mathbf{1}_n \le k \qquad (6)$$
Let $S(k)$ denote the result with parameter k.
2. Select the best cluster in terms of the scan statistic objective among different k:
$$S^* = \arg \max_{S(k), k\in\mathbb{K}} : \eta(S(k)). \qquad (7)$$

Output: the selected connected cluster S^*.

The following lemma describes the efficacy of the procedure above.

Lemma 1. *Let $S_0 = \arg\max_{S\in\Lambda} \eta(S)$ denote the optimal solution. If $|S_0| \in \mathbb{K}$, then $S^* = S_0$.*

Note that now the objective for fixed size is linear in f; since the parameterization only requires an additional linear constraint. The remaining issue is how to characterize the connectivity condition $S \in \Lambda$.

3 Characterizing Exact Connectivity

In Sec.3.1 we first propose an exact characterization of connectivity constraint $S \in \Lambda$ through a linear matrix inequality (LMI) in terms of the binary indicator vector f. We then relax it into a convex SDP constraint in Sec.3.2. In Sec.3.3 we show that the empirical solution of our convex relaxation guarantees connectivity as well as satisfies some nice properties, which then leads to a simple rounding scheme.

3.1 Integer Program Characterization

To deal with connectivity we recall the following lemma from spectral graph theory [16], which elegantly transforms the combinatorial "connectivity" notion into algebraic conditions:

Lemma 2. *Let G be an undirected graph with the unweighted adjacency matrix A and the graph Laplacian matrix L. Then the multiplicity p of the eigenvalue 0 of L equals the number of connected components $C_1, \ldots, C_p$ of the graph.*

We want to guarantee the connectivity of the sub-graph selected by some indicator $f \in \{0,1\}^n$. The next theorem characterizes necessary and sufficient conditions for sub-graph connectivity in terms of an LMI constraint.

Theorem 3. *Given a graph $G = (V, E)$ with unweighted adjacency matrix A, let $S \in V$ be the node set selected by $f \in \{0,1\}^n$. Denote $M = ff'$. Then S forms a connected sub-graph of G if and only if for some positive scalar γ the following LMI holds:*

$$Q(f;\gamma) = Q(M;\gamma) \succeq 0, \qquad (8)$$

where $Q(M) = diag\left((A \circ M - \gamma M)\mathbf{1}_n\right) - A \circ M + \gamma M$.

The proof of Thm.3 involves deriving an expression of the Laplacian matrix L_S of the sub-graph induced by S using the indicator f. Then based on Lemma.2, we apply the Courant-Fischer theorem to characterize the 2nd smallest eigenvalue of L_S. Finally, we apply Finsler's Lemma to convert the condition into an LMI. Details of the proof can be found in supplementary section.

Notice that the LMI constraint Eq.(8) is linear in M. The non-convexity arises from integrality constraint on M which we will relax into convex constraints in the next section. Another point that needs clarification is the role of γ in Eq.(8). This is described in the following corollary. Define $\lambda_2(S)$ to be the second smallest eigenvalue of the Laplacian matrix of S: $\lambda_2(S) = \lambda_2(L_S)$.

Corollary 4. *Let Λ_k be the collection of arbitrarily connected clusters of size k: $\Lambda_k = \{S \subset V : S \text{ is connected}, |S| = k\}$. Let $\lambda_2(\Lambda_k) = \min_{S\in\Lambda_k} : \lambda_2(S)$. M is as defined in Thm.3. Then Λ_k is fully characterized by:*

$$\Lambda_k = \{S \subset V : Q(M;\gamma) \succeq 0, diag(M)'\mathbf{1}_n = k\}, \quad (9)$$

where $\gamma \le \lambda_2(\Lambda_k)/k$.

Remark: (1) Thus solving an integer program with the above constraints on the integer variable M is equivalent to searching for clusters within Λ_k on the

graph. In other words, γ and k parameterize the collection of arbitrarily connected sub-graphs of G. (2) It is well-known that $\lambda_2(S)$ [16] characterizes how well S is connected. γ sets a lower bound on $\lambda_2(S)$. Intuitively larger γ favors "thicker" clusters.

3.2 Convex Relaxation

Note that the symmetric matrix variable M in the connectivity constraint Eq.(8) is binary and has rank one: $M = ff', f \in \{0,1\}^n$. We can relax these non-convex constraints leading to the following convex relaxation to the IP constraint Eq.(8):

$$\begin{aligned} &Q(M;\gamma) \succeq 0 \\ &0 \le M_{ij} \le \min\{M_{ii}, M_{jj}\} \le 1 \\ &M_{ij} - M_{ii} - M_{jj} + 1 \ge 0 \end{aligned} \tag{10}$$

It is easily seen that every binary rank-one matrix M satisfies the above linear constraints. Surprisingly with a small additional constraint the converse is true as well.

Lemma 5. *Let M be any symmetric matrix belonging to the set described by the constraints in Eq.(10). If M is binary, i.e., $M \in \{0,1\}^{n\times n}$, then M has rank-one. Conversely, if M has rank-one with some node having $M_{ii} = 1$, then M is binary.*

3.3 Guarantees on Connectivity & Rounding

We now show that with an additional condition, the support of $diag(M)$ is guaranteed to be connected and satisfies additional properties.

Theorem 6. *Let M be an element of the set described by constraints in Eq.(10) with some diagonal element $M_{ii} = 1$. Let $S = \{i \in V : M_{ii} > 0\}$ be the support of $diag(M)$. Then:*
(a) S forms a connected sub-graph of G.
(b) The induced weighted sub-graph M_S satisfies the following property for any $C \subset S$,

$$\frac{cut\left(C, \bar{C}\right)}{\min\{vol(C), vol(\bar{C}\}} \ge \gamma, \tag{11}$$

where $\bar{C} = S - C$, $cut\left(C, \bar{C}\right) = \sum_{i\in C, j\in\bar{C}, (i,j)\in E} M_{ij}$, $vol(C) = \sum_{i\in C} M_{ii}$.

The proof can be found in supplementary section.

Remark: (1) The additional condition $M_{ii} = 1$ can be imposed when one wants to search for a connected cluster around some particular node i. We observe that by virtue of optimizing an objective function there are usually many diagonal components that achieve this value and so this is usually unnecessary. (2) Note that to generate a sparse support one can also add an L1 penalty on $diag(M)$, or equivalently a size constraint $diag(M)'\mathbf{1}_n \le k$, as will be done in our experiments. (3) If we view M as weights on nodes and edges of G, then (b) shows that empirically γ lower bounds the weighted conductance of S. Larger γ generates thicker clusters.

Typically for any relaxation of a combinatorial optimization problem, some heuristic rounding step is required to convert the continuous solution back to a combinatorially feasible solution. Here we need a rounding scheme to convert $diag(M)$ of Eq.(10) back to an unweighted connected cluster of G. Instead of directly using the support of $diag(M)$, we use an alternative refinement strategy that often leads to a better discrete solution. The next proposition naturally motivates such a scheme.

Proposition 7. *Let M be a symmetric matrix belonging to the set described by constraints in Eq.(10) and S the support of $diag(M)$. Consider any two disjoint connected clusters $C_1, C_2 \in S$. Consider any link (i,j) such that $i \in C_1$, $j \in C_2$, we have,*

$$M_{ii}\,(M_{jj}) \ge \gamma \min\{n_{C_1}, n_{C_2}\}, \tag{12}$$

where n_C denotes the number of 1's in $\{M_{ii} : i \in C\}$.

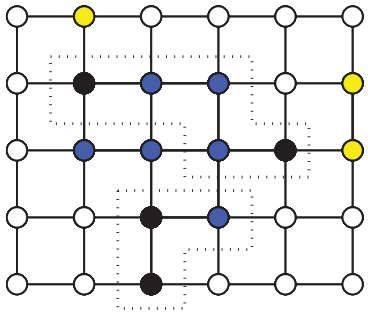


Figure 2: Demonstration of Prop.7.

Remark: Fig.2 depicts the geometric meaning of Prop.7. On a 5×6 grid assume all colored nodes (black,blue,yellow) form the support of $diag(M)$ which is some element described by Eq.(10). Black nodes denote those with $M_{ii} = 1$. Prop.7 guarantees that all blue nodes have at least $M_{ii} \ge \gamma$. In particular, let $C_1 = \{(2,2),(2,3),(2,4),(3,4),(3,5)\}$, $C_2 = \{(4,3),(4,4),(5,3)\}$. The only edge between C_1 and C_2 is ((3,4),(4,4)). By Prop.7 these two nodes will both have $M_{ii} \ge 2\gamma$. On the other hand, yellow nodes may have small M_{ii} values, indicating smaller contribution to the objective $c(M)$.

Intuitively on the support of $diag(M)$, the nodes in the region "spanned" by those with $M_{ii} = 1$ (black and blue) are guaranteed to have large values of M_{ii}.

This suggests that a better scan statistic can often be obtained by discarding those nodes outside of this region. This motivates the following heuristic rounding step for refining $S = supp(diag(M))$ leading to a rounded combinatorial solution to Eq.(1):

Algorithm 2: Rounding
Input: continuous solution $diag(M)$.

1. Let $S = \{i : M_{ii} > 0\}|$ with $L = |S|$. Sort $v \in S$ in descending order: $M_{v_1} \geq \ldots \geq M_{v_L}$.
2. For $l = 1, 2, \ldots, L$, do:
   - Let $V_l = \{v_1, \ldots, v_l\}$. Note that V_l may not be connected.
   - Apply a depth-first search (DFS) from v_1 within V_l to find the connected cluster S_l containing v_1.
3. Among $\{S_l, l = 1, 2, \ldots, L\}$, select the best cluster: $S^* = \arg\max_l : c(S_l)$.

Output: the selected connected cluster S^*.

Remark: Truncating the sorted list of nodes is equivalent to thresholding M_{ii}. By Thm.6 and Prop.7, thresholding at γ guarantees that the depth-first search starting from any node with $M_{ii} = 1$ finds the region spanned by all nodes with $M_{ii} = 1$. Optimizing over thresholds generates larger families of rounded discrete solutions leading to better objective values.

4 Alternative Approaches

Typical scan statistic methods scan parametric shapes such as rectangles, circles in spatial graphs or neighborhood balls in general graphs.

Simulated Annealing: Currently the simulated annealing algorithm [12] is the only algorithm capable of searching for arbitrary shaped connected clusters. This algorithm propagates a region by making heuristic choices based on adding/removing nodes.

An alternative approach is to augment objective function with regularization terms. This imposes smoothness conditions on graph structures by adding graph regularization terms. We experiment with these methods and describe them briefly here.

Edge-lasso regularization: This has been proposed recently by [9] to directly estimate the signal by penalizing an edge-lasso regularization term:

$$\min_{\hat{x}} : \ ||x - \hat{x}||^2 + \lambda ||B\hat{x}||_1, \tag{13}$$

where B is the oriented incidence matrix of $G(V, E)$, defined as: for each undirected edge $e(u, v) \in E$, randomly define an orientation $e^+ = u, e^- = v$ and construct $B \in \{-1, 0, 1\}^{|E|\times|V|}$ with $B_{e,i} = 1$ if $i = e^+$, -1 if $i = e^-$ and 0 otherwise. We denote this method by L1R-a since it penalizes the L1 norm of differences of edges. A variant of this is to augment the edge-lasso penalizing term to our objective Eq.(6):

$$\min_{0\leq f\leq 1} : \ -f'x + \lambda ||B\hat{f}||_1, \quad s.t.\ f'\mathbf{1}_n \leq k \tag{14}$$

We denote this method as L1R-b.

Graph Laplacian regularization: [14] proposes a graph Laplacian regularization method to search for anomalous clusters with small RatioCut values. However, their framework only works when the size of the cluster is completely balanced, i.e. approximately $n/2$. This method in our setting amounts to adding a graph Laplacian regularizing term to Eq.(6):

$$\min_{0\leq f\leq 1} : \ -f'x + \lambda f'Lf, \quad s.t.\ f'\mathbf{1}_n \leq k. \tag{15}$$

We denote this method by L2R.

Notice that none of the above three regularization methods explicitly imposes connectivity. So we apply the same heuristic rounding step (Algorithm 1) to the continuous result ($\hat{x}$ for L1R-a, f for L1R-b and L2R) as described in Sec 4.2 to generate connected clusters.

5 Experiments

In this section we present experiments on both synthetic and real data sets. We compare our exact connectivity (EC) method (Algorithm 2) against scanning with simple families of shapes such as rectangles and neighborhood balls. We then compare against simulated annealing (SA) and graph regularization methods (Sec.4) that do not explicitly parametrize the shapes. For our EC method we vary k and γ and obtain the best solution through the model selection step of Algorithm 2. For rectangle scanning (Rect) on lattice, we enumerate all possible rectangles and choose the region with maximum value of scan statistic. For the neighborhood ball (NB) scanning method on general graphs, we enumerate each node v and scan its k-hop neighborhood balls $N_v(k) = \{u \in V : d(u, v) \leq k\}$ with different values of k, where $d(u, v)$ is the shortest hop distance from u to v. For the simulated annealing algorithm we start the search from a randomly selected node with 50 retries. For the graph regularization methods we vary the trade-off coefficient λ and size parameter k and optimize the scan statistic.

Toy Example: The ground truth on a 48-node grid is shown in Fig.3(a). The input x is noiseless: $x_v = 1$ for red nodes and 0 elsewhere. In this case the global optimal solution with maximum scan statistic (Eq.(2)) is the smallest connected cluster linking two parts, as shown in Fig.3(c). $x_v = 0$ at position (3,5) simulates

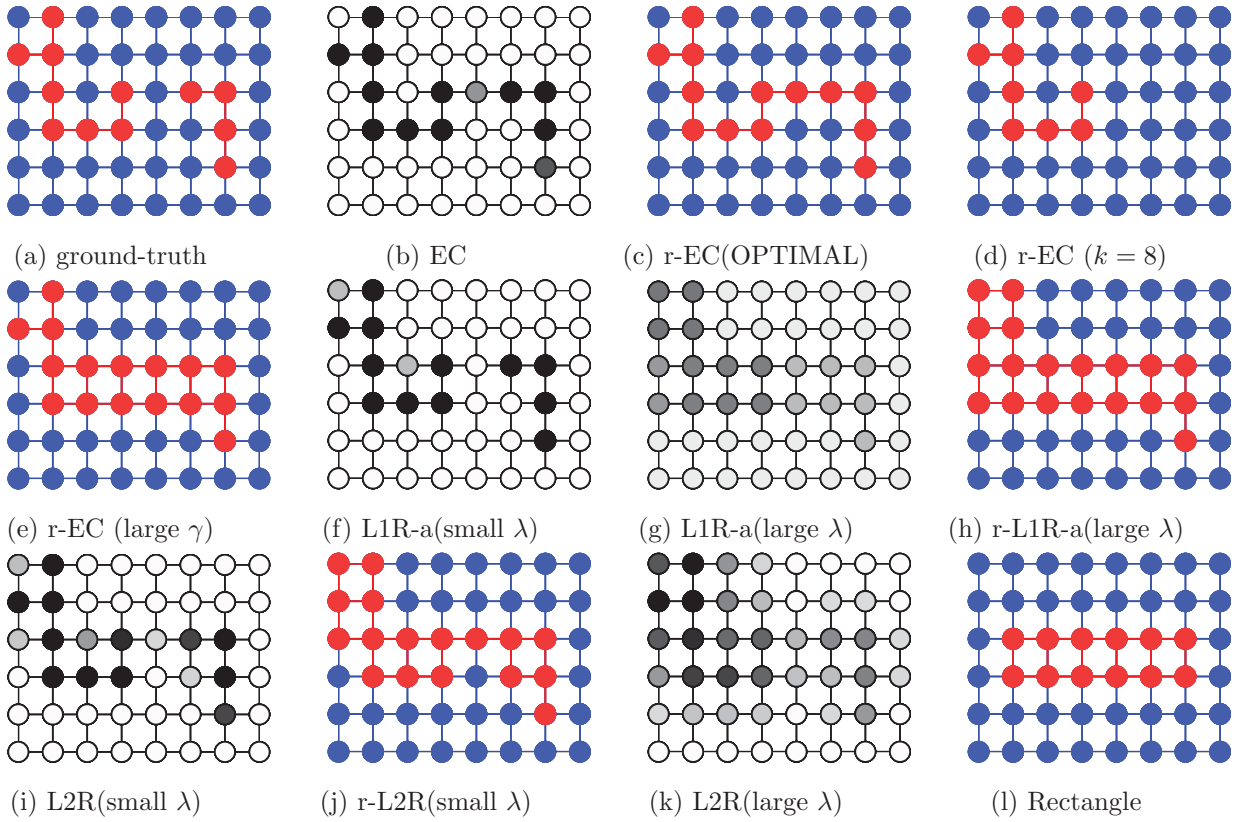


Figure 3: Recovery for toy example by EC, L1R-a, L1R-b, L2R and Rect. Blue-red plots ("r-" means rounded) are discrete results after rounding; grey-scale plots are continuous results. (a) shows ground truth. EC links two separated parts and yields a very clean and sparse continuous result (b), which is then rounded to the GLOBALLY OPTIMAL result of (c). (d) shows recovery of only left part when k is restricted. (e) shows "thick" recovery for larger γ. L1R-a (similarly for L1R-b) with small λ completely ignores connectivity in (f), and is over-regularized with large λ as in (g,h). L2R with small/large λ tend to include irrelevant nodes for thick clusters as in (i-k).

the effect of a node with very poor SNR. It turns out for this example one has to recover both sub-clusters accurately to maximize the scan statistic value.

Various algorithms including rectangle scanning, graph regularizations, L1R-a, L1R-b, L2R, are compared against our EC method. For L1R-a, L1R-b and L2R we solve Eq.(13,14,15) with $k = 13$ and various values of λ before applying the rounding step. For EC we vary k and γ to demonstrate the effect of size-constraints and recovery.

Fig.3(b) shows that EC with $k = 12.5$, $\gamma = 0.005$ generates a clean and sparse continuous solution which effectively links the two disconnected parts and recovers the optimal cluster in (c) with $\eta(S) = 12/\sqrt{13} = 3.33$. Moreover, EC with size restriction $k = 8$ recovers the left part in (d) and recovers thick cluster in (e) with $\gamma = 0.009$. L1R methods with small λ loses connectivity in (f), only recovers the left part of (d) after rounding with $\eta(S) = 2.83$. With large λ L1R methods tend to generate piece-wise constant solutions (g) (similar behavior is observed in [9]), and results in poor recovery (h) with $\eta(S) = 2.75$. L2R with small λ tends to include irrelevant nodes for thick clusters (i) and recovery (j) with $\eta(S) = 2.91$. L2R with large λ generally includes irrelevant nodes (k). All regularization methods fail to recover irregularly shaped clusters.

Synthetic Experiment: We then conducted detection experiments for the positive elevated means model introduced in [3](see Sec.2), on a random geometric graph (RGG) and a stochastic block model (SBM) graph. We randomly generate the graphs (64 nodes) and choose irregular shaped ground-truth anomalous clusters (17 nodes) for both graphs. Due to space limits we only depict the RGG in Fig.4. For the SBM graph we generate two densely connected graphs, G_1, G_2 that are disjoint. We then link nodes of G_1 to nodes of G_2 with some small probability. The anomalous cluster S in this case is chosen to be a connected sub-graph that stretches from one cluster to the other. The idea behind this example is that graph-regularization tend to favor regular shapes, or "fat/thick" clusters. More importantly our algorithm attempts to find irregular shapes such that the conduc-

tance of the sub-graph is approximately larger than γ. On the other hand graph-regularization schemes are based on graph-partitioning and attempt to find partitions that have low conductance. Consequently, for a stochastic block model they tend to favor choosing nodes only from one of the clusters rather than both.

We carry out 300 null/alternative tests respectively, with fixed noise level $\sigma = 1$ and different values of signal strength μ (Eq.(3)). We illustrate performance with respect to the normalized SNR: $\Lambda = \mu\sqrt{|S|}/\sigma$ [3]. For each test we randomly generate Gaussian noise and apply various methods including Rect, NB, L1R-a, L1R-b, L2R and EC followed by the heuristic rounding step, while being agnostic to the size, shape or position of the ground-truth S. For L1R-a, L1R-b, L2R and EC we try a range of values for $\lambda, k, \alpha, \gamma$ and select the best discrete result. $\gamma \in \{0.001, 0.002, 0.004, ..., 0.064\}$ for EC. $k \in \{6, 9, 12, 15, 18, 21, 24, 27\}$ for L1R-b, L2R and EC. $\lambda \in \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1\}$ for L1R-a, L1R-b and L2R.

For each SNR setting we then threshold the scan statistic values of various tests with different thresholds for decision of H_0/H_1. We then compare decision results against ground-truth and tabulate AUC performance in Tab.1. We also tabulate the detection performance of various methods with FA rate at approximately 10% in Tab.2 for both RGG and SBM. As argued before EC is superior on SBM as regularization favors picking nodes from one of the clusters. Both tables demonstrate that our method compares favorably to SA and significantly outperforms other methods

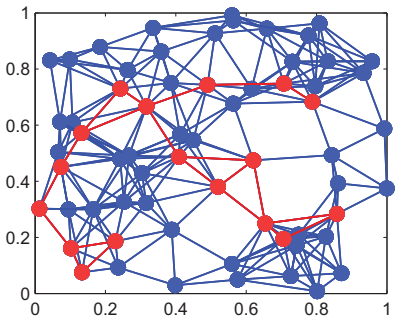


Figure 4: Random Geometric Graph Used in Experiment

Recovery for Disease Outbreak Dataset: We apply our framework for the setting of disease outbreak detection as in [13]. We use real population data from the northeastern U.S. geographic counties (129 nodes), including Massachusetts, New York, Vermont, Maine, New Hampshire, Connecticut and Rhode Island in Fig.5(a,b). The ground truth reveals disease outbreak in two regions: Lake Ontaria coast (left part in (a)) and Hudson River region (right part in (a)).

We consider the problem of outbreak detection, where

Table 1: AUC performance of various algorithms with different normalized SNR $\Lambda = \mu\sqrt{|S|}/\sigma$ on RGG.

AUC		normalized SNR		
		3.0	3.5	4.5
RGG	EC	0.8639	0.9158	0.9732
	SA	0.8623	0.9123	0.9758
	Rect	0.7699	0.8040	0.9195
	L1R-a	0.8314	0.8892	0.9640
	L1R-b	0.7912	0.8437	0.9561
	L2R	0.8481	0.8908	0.9619

Table 2: Detection rate performance of various methods at false alarm rate of 10%. SA performs similar to EC and is not shown here. EC significantly outperforms other methods. RGG is a random geometric graph. SBM refers to a stochastic block model.

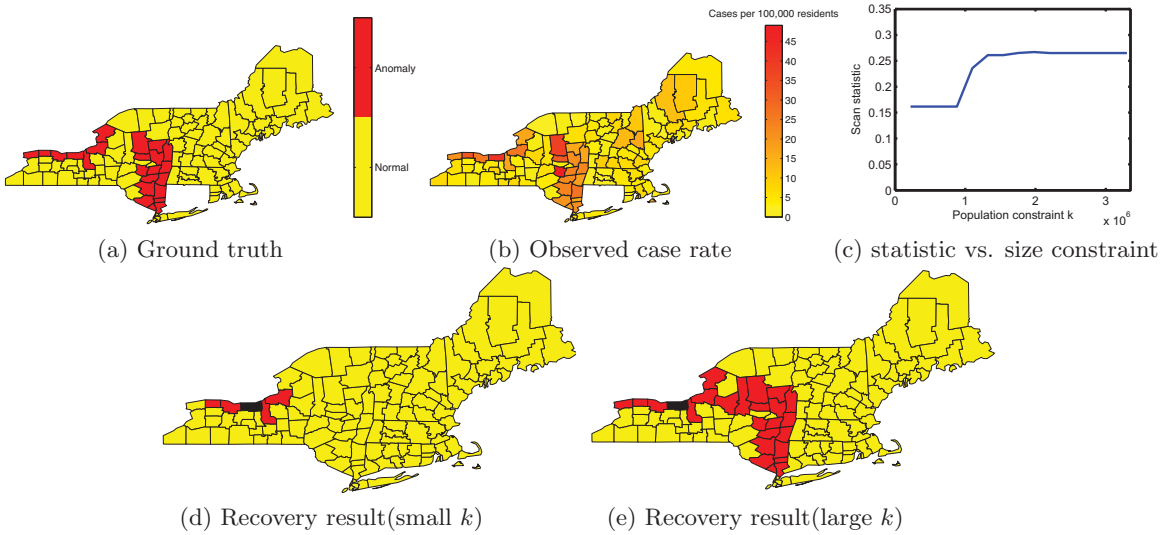
AUC		normalized SNR		
		3.0	3.5	4.5
RGG	EC	63.9%	79.4%	94.2%
	Rect	52.1%	60.2%	79.6%
	L1R-a	60.1%	77.8%	91.5%
	L1R-b	53.7%	67.3%	82.5%
	L2R	55.6%	75.9%	91.8%
SBM	EC	58.6%	77.8%	91.2%
	Rect	33.5%	50.2%	73.8%
	L1R-a	53.9%	72.9%	88.6%
	L1R-b	43.5%	61.6%	85.2%
	L2R	52.0%	71.9%	87.9%

the clusters consist of adjacent counties forming connected sub-graphs, under the Poisson model. The number of disease cases c_i within county i is a Poisson random variable with parameter $N_i\lambda_i$, where N_i is the population of county i and $\lambda_i = \mu_0$ for normal counties and $\lambda_i = \mu_1 > \mu_0$ for anomalous counties. μ_0 is assumed to be known, which in reality can be estimated by the average rate over years. μ_1 is unknown. We are interested in distinguishing between the null hypothesis $H_0 : c_i \sim Poisson(N_i\mu_0), \forall i$ and the alternative $H_1 = \bigcup_{S\in\Lambda} H_{1,S}$, where $H_{1,S} : c_i \sim Poisson(N_i\mu_1), \forall i \in S; c_i \sim Poisson(N_i\mu_0), \forall i \notin S$. [3] suggests that the following statistic performs well for rejecting H_0 for large values:

$$G(S) = \sqrt{\sum_{i\in S} N_i}\left(\left(\frac{\sum_{i\in S} c_i}{\sum_{i\in S} N_i}\right) - \mu_0\right) \tag{16}$$

$$\Rightarrow G(f) = \left(\frac{f'C}{\sqrt{f'N}} - \mu_0\sqrt{f'N}\right), \tag{17}$$

where N and C are population and case vectors across the counties and f is indicator of S. While non-convex in f, $G(f)$ is monotonic in $f'N$; this enables us to apply exactly the same convexifying trick on $G(f)$ as



(a) Ground truth (b) Observed case rate (c) statistic vs. size constraint

(d) Recovery result(small k) (e) Recovery result(large k)

Figure 5: (a) shows the county map of northeastern U.S. including 7 states, with ground-truth clusters corresponding to Lake Ontaria Coast (left) and Hudson River region (right). (b) shows the observed case/population rates of each county. (c) plots the scan statistic G vs. population constraint parameter k, which has two flat parts, the lower shown in (d) and the higher in (e). We set $M_{ii} = 1$ of the black county which has the highest case/population rate, indicating we want to search for connected regions around this county.

described in Sec. 2.1 and convert this into an equivalent 2-step procedure. Thus we can apply the same Algorithm 2, with the following modified sub-routine:

$$\begin{aligned} \max: \quad & diag(M)'C \\ s.t. \quad & Q(M;\gamma) \succeq 0 \\ & 0 \le M_{ij} \le \min\{M_{ii}, M_{jj}\} \le 1 \\ & M_{ij} - M_{ii} - M_{jj} + 1 \ge 0 \\ & diag(M)'N \le k \end{aligned} \tag{18}$$

and the modified selection criterion $c(S) = G(S)$ in step 2,3 of Algorithm 2.

For simulation a benchmark dataset (numbers of disease cases for both H_0 and H_1) is first constructed using real data population. Specifically, we generate the numbers of disease cases c_i in each county i according to Poisson distribution with parameter $N_i\lambda_i$, where $\lambda_i = \mu_0 = 5 \times 10^{-5}$ for normal counties and $\lambda_i = \mu_1 = 4\mu_0$ for anomalous counties. Fig.5(b) shows the empirical case/population rates of each county. We then apply our algorithm to detect the outbreaks. By Thm.6 our method needs a seed node. For this we pick the county with the largest incidence rate, which is colored black in (d,e) as shown. We then search for connected regions around this most severe county. We plot the scan statistic G against the population threshold parameter k of Eq.(18) in (c). This curve has two flat regions, with the lower one corresponding to Lake Ontaria coast in (d), and the higher one corresponding to the globally optimal cluster in (e) which links Lake Ontario coast with Hudson River region.

Discussion:

(1) Our method finds irregularly-shaped connected clusters as is claimed in Thm.6. Even when the optimal cluster consists of two disconnected clusters, by Prop.7 our algorithm is able to select the two counties linking Lake Ontario coast with Hudson River region, yielding the globally optimal result (e).

(2) By restricting the size, multiple clusters are identified as seen in the statistic-size plot. Our method allows estimating multiple outbreak regions with different sizes.

(3) Other alternative methods have various drawbacks. SA only realizes the large cluster and is not sufficiently flexible to deal with multiple outbreak regions. In addition our recovery results also appear to be sparse and clean in comparison to any other regularization method (which are not presented here due to lack of space), which typically contain large number of false alarms (i.e. counties that are not part of the outbreak).

(4) Our method can deal with up to 300 nodes using sedumi/cvx under matlab environment. Computation complexity for solving SDP problems has been a barrier for many machine learning algorithms [17, 18]. For scalability to larger graphs, a divide-and-conquer strategy can be applied; our approach can be used in 2nd stage for locally refined search.

Acknowledgements

This work was partially supported by NSF Grant CCF-1320547 and U.S. Department of Homeland Security under Award Number 2008-ST-061-ED0001.

References

[1] G. P. Patil and C. Taillie. Geographic and network surveillance via scan statistics for critical area detection. In *Statistical Science*, volume 18, pages 457–465, 2003.

[2] L. Pickle M. Kulldorff, L. Huang and L. Duczmal. An elliptic spatial scan statistic. In *Statistics in Medicine*, volume 25, 2006.

[3] E. Arias-Castro, E. J. Candes, and A. Durand. Detection of an anomalous cluster in a network. In *The Annals of Statistics*, volume 39, pages 278–304, 2011.

[4] D. J. Marchette C. E. Priebe, J. M. Conroy and Y. Park. Scan statistics on enron graphs. In *Computational and Mathematical Organization Theory*, 2006.

[5] L. Devroye Addario-Berry, N. Broutin and G. Lugosi. On combinatorial testing problems. In *The Annals of Statistics*, volume 38, pages 3063–3092, 2010.

[6] D. Donoho E. Arias-Castro and X. Huo. Near-optimal detection of geometric objects by fast multiscale methods. In *IEEE Transactions on Information Theory*, volume 51, pages 2402–2425, 2005.

[7] H. Helgason E. Arias-Castro, E. J. Candes and O. Zeitouni. Searching for a trail of evidence in a maze. In *The Annals of Statistics*, volume 36, pages 1726–1757, 2008.

[8] R. Nowak A. Singh and R. Calderbank. Detecting weak but hierarchically-structured patterns in networks. In *AISTATS*, 2010.

[9] A. Rinaldo J. Sharpnack and A. Singh. Sparsistency of the edge lasso over graphs. In *AISTATS*, volume 22, pages 1028–1036, 2012.

[10] J. Naus J. Glaz and S. Wallenstein. *Scan Statistics*. Springer, New York, 2001.

[11] D. J. Marchette and C. E. Priebe. Scan statistics for interstate alliance graphs. In *Connections*, volume 28, pages 43–64, 2008.

[12] L. Duczmal and R. Assuncao. A timulated annealing strategy for the detection of arbitrarily shaped spatial clusters. In *Computational Statistics and Data Analysis*, volume 45, pages 269–286, 2004.

[13] M. Kulldorff L. Duczmal and L. Huang. Evaluation of spatial scan statistics for irregularly shaped clusters. In *Journal of Computational and Graphical Statistics*, volume 15, pages 428–442, 2006.

[14] A. Rinaldo J. Sharpnack and A. Singh. Changepoint detection over graphs with the spectral scan statistic. In *arXiv: 1206.0773v1*, 2012.

[15] D. S. Johnson, M. Minkoff, and S. Phillips. The prize collecting steiner tree problem: theory and practice. In *ACM-SIAM Symp. on Discrete Algorithms*, 2000.

[16] F. Chung. *Spectral graph theory*. American Mathematical Society, 1996.

[17] Z. Xu and R. Jin. Efficient convex relaxation for transductive support vector machine. In *NIPS*, 2007.

[18] N. Vasconcelos A. B. Chan and G. R. G. Lanckriet. Direct convex relaxations of sparse svm. In *ICML*, 2007.