ANOMALOUS CLUSTER DETECTION

Jing Qian *Venkatesh Saligrama* *Yuting Chen*

Boston University
Boston, MA, 02215

ABSTRACT

We consider the problem of anomalous cluster detection (ACD) on a graph under the elevated mean Gaussian model, where each node is associated with a feature. Under the null hypothesis, features are i.i.d. standard Gaussian, while under the alternative, there is an unknown connected cluster of nodes whose features are i.i.d. Gaussian with positive mean and unit variance instead. For this problem the GLRT scan statistic is usually adopted; however there are very few practical algorithms that target arbitrarily connected clusters. We formulate this problem as an integer program (IP) in terms of indicator variables, and characterize the connectivity of a cluster by a linear matrix inequality (LMI) constraint. We then propose a convex relaxation of the IP together with a rounding scheme, leading to a completely convex formulation for computing the scan statistic over arbitrarily connected clusters. Synthetic and real experiments justify our idea.

Index Terms— Anomalous Cluster Detection, Connectivity, Semi-Definite Programming

1. INTRODUCTION

Anomalous cluster detection (ACD) on a graph refers to the problem of detecting whether or not there is a connected cluster of nodes that behave differently from the rest nodes of the graph. Such a problem has been extensively studied [1, 2, 3, 4, 5, 6]. Usually ACD is formulated as a hypothesis testing problem embedded in a graph $G = (V, E)$, where each node i is equipped with an independent random variable x_i. In this paper we focus on the elevated mean Gaussian model, where the goal is to distinguish between the null hypothesis that observations over all nodes come from i.i.d. standard normal distribution: $x_i \sim N(0,1), \forall i \in V$, against the alternative that observations of nodes on an unknown connected subgraph, $S \subset V$, involve some signal strength: $x_i \sim N(\mu, 1), \forall i \in S$, for $\mu > 0$. [4] has shown that under some conditions the test of rejecting the null hypothesis for large values of the following scan statistic is statistically optimal or near-optimal:

$$\max_{S \in \Lambda} : \ \eta(S) = \sum_{i \in S} x_i / \sqrt{|S|} \qquad (1)$$

where $\Lambda = \{S \subseteq V : S \text{ is connected}\}$, and f is the indicator of S, i.e. $f_i = 1$ for $i \in S$ and 0 otherwise.

While existing work mainly focuses on statistical decision aspects of the problem, they usually consider relatively simple graph structures, and only scan simple clusters including rectangles, circles or nearest-neighbor balls [7, 8, 9]. Our motivation of searching for arbitrarily connected clusters come from real applications. For example, consider the problem of disease outbreak detection [10, 11] shown in Fig.1. A disease outbreak could happen around a river, leading to elevated numbers of disease cases in those spatially adjacent counties near the river, which form an irregular connected cluster within the graph representation. Other problems such as surveillance and network intrusion can also be cast in this manner. However, for arbitrary shapes the simulated annealing method

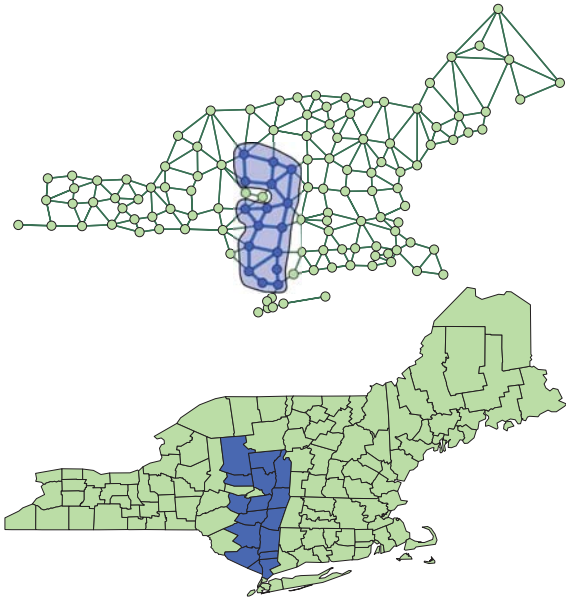


Fig. 1. County map of northeast U.S. and its graph representation. Shaded counties in the lower panel, which refer to Hudson River region, represent a possible disease outbreak. The nodes corresponding to these counties form an irregularly shaped connected cluster.

[12, 11] seems to be the only viable algorithm, which requires multiple restarts and often many iterations to converge. The spectral scan statistic method [13] based on graph regulariza-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tion does not restrict cluster shapes, but it can not guarantee connectivity. Moreover this method favors balanced partitions with small conductance, which may not be the case as in Fig.1. In contrast, our method allows for arbitrary connected subgraphs with explicit control on size.

The main contribution of this paper is to present a convex program for computing the scan statistic Eq.(1) over arbitrarily connected clusters. We first formulate it as an integer program (IP), and characterize the connectivity constraint $S \in \Lambda$ in terms of a linear matrix inequality (LMI) constraint. We then propose a convex relaxation to the IP problem, together with a novel rounding scheme that refines for a better combinatorial solution. Synthetic and real experiments demonstrate the efficacy of our idea.

2. ANOMALOUS CLUSTER DETECTION UNDER ELEVATED MEAN GAUSSIAN MODEL

Let $G = (V, E)$ be an undirected graph with n nodes $|V| = n$, the unweighted adjacency matrix A and unnormalized Laplacian matrix L. Let $S \subset V$ be indicated by $f \in \{0,1\}^n$. ACD under the elevated-mean Gaussian model can be formulated into a composite hypothesis testing problem. Specifically, the observation x_i of node i follows standard normal distribution under the null hypothesis $H_0 : x_i \sim N(0,1), \forall i \in V$. The alternative hypothesis is $H_1 = \bigcup_{S\in\Lambda} H_{1,S}$, and each $H_{1,S}$ is parameterized by S: $H_{1,S} : x_i \sim N(\mu, 1), \forall i \in S$; $x_i \sim N(0,1), \forall i \notin S$, where $\mu > 0$ is signal strength, and S is some unknown connected anomalous cluster: $S \in \Lambda$.

Let $S(f)$ denote the subgraph indicated by f. The scan statistic Eq.(1) can be rewritten in terms of f as:

$$\max_{f\in\{0,1\}^n, S(f)\in\Lambda} : \eta(f) = \frac{f'x}{\sqrt{f'\mathbf{1}_n}}. \tag{2}$$

Note that $\eta(f)$ is not concave in f, which makes the objective non-convex, not to mention the binary nature of f. We propose to convexify the objective by transforming the problem into a 2-step procedure, which involves first solving a family of sub-problems parameterized by size of S, followed by a model selection step.

Algorithm 1: Computing Scan Statistic (IP)

Input: n observations $\{x_1, \ldots, x_n\}$ of the nodes, adjacency matrix A, size parameter set K.

1. For different values of $k \in K$, solve:

$$\max_{f\in\{0,1\}^n, S(f)\in\Lambda} : f'x \quad s.t.\ f'\mathbf{1}_n \le k \tag{3}$$

Let $S(k)$ denote the obtained cluster with parameter k.

2. Select the best cluster in terms of $\eta(S)$ over various k:

$$S^* = \arg\max_{S(k), k\in K} : \eta(S(k)). \tag{4}$$

Output: the selected connected cluster S^*.

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

When the size parameter set K is rich enough, the above procedure is equivalent to the original Eq.(2). In Sec.3 we provide a convex characterization of the connectivity constraint $S \in \Lambda$, which leads to a convex relaxed SDP problem.

3. CONNECTIVITY & CONVEX RELAXATION

In this section we characterize connectivity of S, and provide a convex relaxation to the IP subroutine Eq.(3).

3.1. Characterizing Connectivity

Our main theorem characterizes the necessary and sufficient condition for sub-graph connectivity.

Theorem 2. *Given $G = (V, E)$, let $S(f) \in V$ be the node set selected by $f \in \{0,1\}^n$. Denote $F = ff'$. Then S forms a connected cluster if and only if for some positive scalar γ,*

$$Q(f;\gamma) = Q(F;\gamma) \succeq 0, \tag{5}$$

where $Q(F;\gamma) = diag\left((A \circ F - \gamma F)\mathbf{1}_n\right) - A \circ F + \gamma F$, $\circ$ denotes entry-wise matrix multiplication, and $Q \succeq 0$ denotes Q is positive semi-definite.

We sketch the proof here. We first "select" the induced adjacency matrix A_S, thus the Laplacian L_S of S using F. Courant-Fischer theorem is applied to characterize the 2nd smallest eigenvalue $\lambda_2(S) = \lambda_2(L_S)$. By spectral graph theory [14] S is connected if and only if $\lambda_2(S) > 0$. Applying Finsler's Lemma then converts the condition into an LMI. Details are omitted due to lack of space.

Eq.(5) is in terms of $F = ff'$. We can use the equivalent $f = diag(F)$ to replace f in the objective of Eq.(3). The next corollary shows that γ in Eq.(5) and the size parameter k parameterize the collection of all connected sub-graphs Λ.

Corollary 3. *Let $\Lambda_k = \{S \in \Lambda : |S| = k\}$ be the set of connected clusters of size k. Let $\lambda_2(\Lambda_k) = \min_{S\in\Lambda_k} : \lambda_2(S)$. F is defined in Thm.2. Then Λ_k is fully characterized by:*

$$\Lambda_k = \{S \subset V : Q(F;\gamma) \succeq 0, diag(F)'\mathbf{1}_n = k\}, \tag{6}$$

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

Remark: (1) Each feasible integer variable F satisfying Eq.(6) corresponds to a connected cluster in Λ_k, and vice versa. Solving an IP problem with constraints in Eq.(6) is equivalent to searching over $S \in \Lambda_k$. (2) It is well-known that $\lambda_2(S)$ [14] is related to the conductance of S which characterizes how well S is connected. Intuitively this implies that setting a larger γ restricts the search to only thicker clusters, while small γ allows irregular thin shapes.

3.2. Convex Relaxation & Rounding Scheme

Notice that the variable matrix F in Eq.(5) is binary and has rank one: $F = ff', f \in \{0,1\}^n$. We propose the following linear relaxation to the integer variable F:

$$\begin{aligned} &0 \le F_{ij} = F_{ji} \le F_{ii} \le 1, \ \forall i \ne j \\ &F_{ij} - F_{ii} - F_{jj} + 1 \ge 0, \ \forall i \ne j \end{aligned} \tag{7}$$

It is obvious that every binary rank-one matrix satisfies Eq.(7). On the other hand, the first constraint enforces $F_{ij} = 0$ if either node i or node j is not selected. When both are selected ($F_{ii} = F_{jj} = 1$), the second inequality ensures $F_{ij} = 1$ to approximate indicator matrix $F = ff'$.

Next we present a rounding scheme to convert the continuous solution $diag(F)$ back to a combinatorially feasible solution, i.e. a connected cluster $S \in \Lambda$.

Algorithm 2: Rounding

Input: continuous solution $diag(F)$.

1. Let $S = \{i : F_{ii} > 0\}|$ with $L = |S|$. Sort $v \in S$ in descending order: $F_{v_1 v_1} \ge \ldots \ge F_{v_L v_L}$.
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: $S^* = \arg\max_l : \eta(S_l)$.

Output: the selected connected cluster S^*.

The intuition is that while keeping connectivity, it can lead to better objective to remove those nodes with small values of F_{ii}, which have less contribution to the objective.

The complete algorithm is outlined below.

Algorithm 3: Computing Scan Statistic (convex program)

Input: n observations $\{x_1, \ldots, x_n\}$ of the nodes, adjacency matrix A, size parameter set K.

1. For different values of $k \in K$:
   - Solve the following SDP problem:

$$\begin{aligned} \max: \quad & diag(F)'x \\ s.t. \quad & Q(F;\gamma) \succeq 0 \\ & 0 \le F_{ij} = F_{ji} \le F_{ii} \le 1, \ \forall i \ne j \\ & F_{ij} - F_{ii} - F_{jj} + 1 \ge 0, \ \forall i \ne j \\ & diag(F)'\mathbf{1}_n \le k \end{aligned} \tag{8}$$

   - Apply rounding (Algorithm 2) on the solution $F(k)$ and obtain the connected cluster $S(k)$.
2. Select the best cluster in terms of $\eta(S)$ over various k:

$$S^* = \arg \max_{S(k), k \in K} : \eta(S(k)). \tag{9}$$

Output: the selected connected cluster S^*.

4. EXPERIMENTS

In this section we present experiments on both synthetic and real data sets. We compare our exact connectivity (EC) method (Algorithm 3) against scanning rectangles (Rect) [9] and several other approaches listed below.

Alternative Approaches:

The simulated annealing [12] is the only algorithm capable of searching for arbitrarily connected clusters. It propagates a region by heuristically adding/removing one node at each ietaration. We denote this algorithm as SA.

[6] proposes to directly estimate signal strength by penalizing an edge-lasso regularization term:

$$\min_{\hat{x}} : \ ||x - \hat{x}||^2 + \lambda ||B\hat{x}||_1 \tag{10}$$

where B is the oriented incidence matrix. Details can be found in [6]. We denote this method by L1R-a.

A variant is to augment the edge-lasso penalizing term to our objective, which we denote by L1R-b:

$$\min_{0 \le f \le 1} : \ -f'x + \lambda ||B\hat{f}||_1, \quad s.t. \ f'\mathbf{1}_n \le k \tag{11}$$

[13] proposes a graph Laplacian regularization method to search for anomalous clusters with small RatioCut values. However, their method only works when the cluster is completely balanced, i.e. approximately of size $n/2$. A similar method in our setting, which is denoted by L2R, amounts to:

$$\min_{0 \le f \le 1} : \ -f'x + \lambda f'Lf, \quad s.t. \ f'\mathbf{1}_n \le k \tag{12}$$

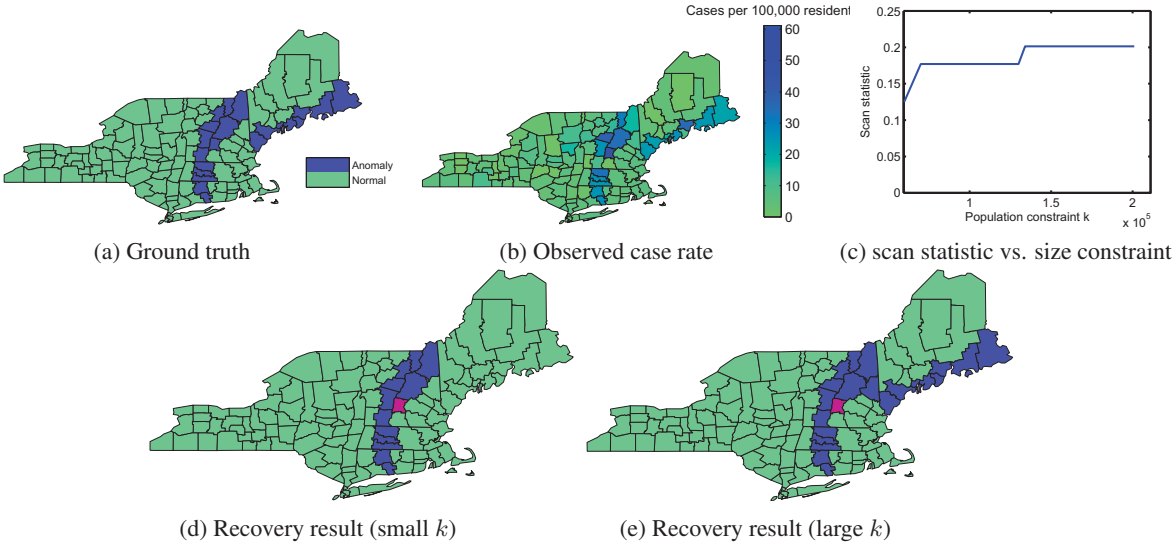
Notice that none of above regularization methods explicitly imposes connectivity. So we apply the same heuristic rounding step (Algorithm 2) to the continuous result ($\hat{x}$ for L1R-a, f for L1R-b and L2R) to generate connected clusters. We vary parameters γ, k and λ to obtain the best solution through the model selection step of Algorithm 3.

Synthetic Detection Experiment:

We conduct ACD experiments on a 8×10 lattice with an irregularly shaped anomalous cluster (12 nodes) shown in Fig.2. 200 null/alternative tests are carried out respectively, with Gaussian noise level $\sigma = 1$ and different values of signal strength μ. We threshold the resulting scan statistic values using different thresholds to generate ROC curves and compute AUC. We illustrate AUC against normalized SNR: $\mu\sqrt{|S|}/\sigma$ in Tab.1. Our EC performs as well as SA, which is roughly enumerating and can be viewed as optimal, and significantly outperforms all other methods.

Recovery for Disease Outbreak Dataset:

We apply our framework for the setting of disease outbreak detection as in [11]. We use real population data from geographic counties (129 counties) of northeastern U.S., including Massachusetts, New York, Vermont, Maine, New Hampshire, Connecticut and Rhode Island shown in Fig.4(a). The



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

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

Fig. 3. (a) shows the county map of northeastern U.S. including 7 states, with ground-truth clusters corresponding to Connecticut River region (left) and New England coast (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 $F_{ii} = 1$ of the pink county which has the highest case/population rate, indicating we want to search for connected regions around this county.

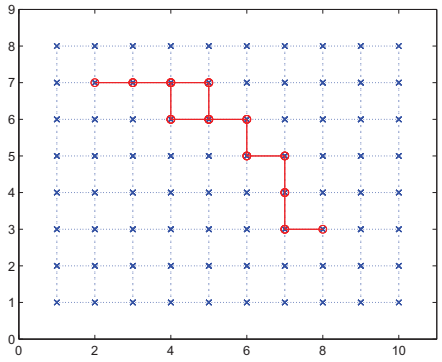


Fig. 2. 80-node lattice with the ground-truth anomalous cluster.

ground truth reveals disease outbreak in two areas: Connecticut River region (left part) and New England coast (right part).

We consider ACD under Poisson model where the anomalous cluster consist of counties that have higher Poisson rate than normal counties. [4] establishes similar scan statistic to Eq.(1) that performs well. Let N and C be the population and disease case vector respectively. We can apply the same Algorithm 3 with modifications for Eq.(8) on objective: $\max : diag(F)'C$, and the size constraint: $diag(F)'N \leq k$.

For simulation we generate numbers of cases C_i in each county from Poisson distribution with parameter $N_i\lambda_i$, where $\lambda_i = 5 \times 10^{-5}$ for normal counties and $\lambda_i = 2 \times 10^{-4}$ for anomalous counties. Fig.3(b) shows the empirical case/population rates. We then apply EC to detect the outbreaks. We plot the scan statistic G against the population threshold k in (c). This curve has two flat regions, the lower corresponding to Connecticut River region (d), and the higher corresponding to the globally optimal cluster (e) which links Connecticut River region with New England coast.

Remark: (1) Our method is able to find irregularly-shaped connected clusters. Furthermore, by constraining the size, multiple clusters are identified from the statistic-size plot. (2) SA only recovers the large cluster similar to Fig.3(e) and is not sufficiently flexible to deal with multiple outbreak regions. In addition our recovery results also appear to be sparse and clean compared to other regularization methods (omitted due to lack of space), which favor thick shapes and typically contain large number of false alarms (i.e. counties that are not part of the outbreak). (3) Currently our method can deal with up to 300 nodes, due to computational barriers in solving SDP. For large graphs one can perform a coarse search before applying our method on the local patch for a refined search. Future directions include developing fast alternating algorithms for solving the subroutine Eq.(8).

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

AUC	normalized SNR			
	3	3.5	4.5	5.5
EC	**0.8787**	**0.9188**	**0.9641**	**0.9924**
SA	0.8679	0.9077	0.9574	0.9763
Rect	0.8125	0.8573	0.9282	0.9658
L1R-a	0.8639	0.9037	0.9640	0.9877
L1R-b	0.8259	0.8738	0.9405	0.9786
L2R	0.8610	0.9058	0.9610	0.9805

5. REFERENCES

[1] Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi, "On combinatorial testing problems," in *The Annals of Statistics*, 2010, vol. 38, pp. 3063–3092.

[2] E. Arias-Castro, D. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," in *IEEE Transactions on Information Theory*, 2005, vol. 51, pp. 2402–2425.

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

[4] E. Arias-Castro, E. J. Candes, and A. Durand, "Detection of an anomalous cluster in a network," in *The Annals of Statistics*, 2011, vol. 39, pp. 278–304.

[5] A. Singh, R. Nowak, and R. Calderbank, "Detecting weak but hierarchically-structured patterns in networks," in *Artificial Intelligence and Statistics*, 2010.

[6] J. Sharpnack, A. Rinaldo, and A. Singh, "Sparsistency of the edge lasso over graphs," in *Artificial Intelligence and Statistics*, 2012, vol. 22, pp. 1028–1036.

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

[8] M. Kulldorff, L. Huang, L. Pickle, and L. Duczmal, "An elliptic spatial scan statistic," in *Statistics in Medicine*, 2006, vol. 25.

[9] D. J. Marchette and C. E. Priebe, "Scan statistics for interstate alliance graphs," in *Connections*, 2008, vol. 28, pp. 43–64.

[10] G. P. Patil and C. Taillie, "Geographic and network surveillance via scan statistics for critical area detection," in *Statistical Science*, 2003, vol. 18, pp. 457–465.

[11] L. Duczmal, M. Kulldorff, and L. Huang, "Evaluation of spatial scan statistics for irregularly shaped clusters," in *Journal of Computational and Graphical Statistics*, 2006, vol. 15, pp. 428–442.

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

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

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